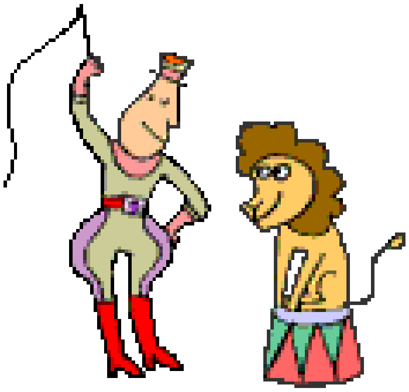
Taming Statistics with TamStat
User Guide

Stephen M. Mansour, PhD



"There's no Lyin' in Statistics"

www.tamstat.com

Note:

There are two versions of TamStat: The desktop application TamStat.exe, and the HTML version. The desktop application is easier to install, but may be limited to the ASCII character set. The HTML version can use the complete Dyalog APL character set. The examples in this book may use non-ASCII characters, in particular the assignment arrow: (←).
For the desktop app only, users may make ASCII substitutions; e.g. for the expression, `A ← 2 + 3`, enter the following instead: `A <- 2+ 3`. Other ASCII substitutions are listed in Appendix A.

All examples assume that `⎕IO←1` and `⎕ML←1`.

ASCII Substitutions:

Function/ Operation	TamStat Workspace and HTML Version	TamStat APP (TamStat.exe) when set to Standard
Assignment	A ← expr	A <- expr
Not equals	A≠B	A <> B or A ne B
Less than or equal to	A≤B	A <= B or A le B
Greater than or equal to	A≥B	A >= B or A le B
Negative numbers	¯5 3 ¯1	-5 3 -1
Multiplication	A × B	A * B
Division	A ÷ B	A / B
Power	A * B	A ^ B

Acknowledgements

I wish to thank Jerry M. Brennan, PhD, Michael Baas, Joshua David, and Bill Koko for their assistance in developing TamStat.

I would also like to thank my wife, Gina, for her patience and support during this endeavor.

Cover art by the author

Taming Statistics with TamStat

Contents

Chapter 0 – A Few Preliminaries....7
0.1 Installation....7
0.1.1 Windows Installation....7
0.1.2 Mac Installation....7
0.1.3 Linux Installation....7
0.2 APL Syntax....8
0.2.1 Data and Variables....8
0.2.2 Functions....8
0.2.3 Order of operations....9
0.2.4 Operators....10
0.2.5 Namespaces....11
0.5 Exercises....12
Chapter 1 - Descriptive Statistics....13
1.1 Importing and Editing Databases in TamStat....15
1.1.2 Desktop Version....15
1.1.2 HTML Version....17
1.1 Tables....19
1.1.1 Categorical Data – frequency, relative frequency....19
1.1.2 Contingency tables....20
1.1.3 Numeric Data – frequency, cumulative frequency, binning....22
1.2 Summary Statistics....24
1.2.1 Measures of Quantity....24
1.2.1 Measures of Center....24
1.2.2 Measures of spread....25
1.2.3 Measures of position....27
1.2.4 Measures of shape....27
1.2.5 Measures of Association....28
1.2.6 Summary Function Wizard....30
1.3 Graphics....32
1.3.1 Pie Charts....32
1.3.2 Bar Charts....32
1.3.3 Histograms....34
1.3.4 Stem-and-Leaf Displays....35

1.3.5 Box Plots 36
1.3.7 Scatter Plots 37
1.3.6 Dot Plots 38
1.3.7 Graphics Wizard 38
1.4 Obtaining Summary Statistics from tabular data 40
1.6 Exercises 42
Chapter 2 – Probability 43
2.1 Theoretical Probability 43
2.2 Empirical Probability 44
2.3 Rules of Probability 44
2.3.1 The Complement Rule 44
2.3.2 The Multiplication Rule 45
2.3.3 The Addition Rule 46
2.3.3 The Conditional Rule 48
2.3.4 Probability Wizard 48
2.4 Bayesian Statistics 52
2.4.1 Bayesian Wizard 53
2.5 Counting Rules 54
2.5.1 Multi-Step Experiment 54
2.5.2 Factorials, Permutations and Combinations 54
2.6 Random Variables and Probability Distributions 55
2.6.1 Discrete Distributions 60
2.6.2 Continuous Distributions 62
2.6.2 Scalar Characteristics of Distribution Functions 67
2.6.3 Distribution Wizard 69
2.6.4 Statistical Tables 75
2.7 Exercises 78
Chapter 3 - Inferential Statistics 79
3.1 Sampling Distributions 79
3.2 Confidence Intervals 80
3.2.1 Means 80
3.2.2 Proportions 82
3.2.3 Confidence Intervals Using the Wizard 84
3.2.4 Sample Size Wizard 86
3.3 Hypothesis Tests 87
3.3.1 Means 87
3.3.2 Proportions 88
3.4 Two Sample Tests 89
3.4.1 Difference of Means (Independent Samples) 89

3.3.5 Mean of the Differences (Paired Data) 90
3.5 Inferences about Population Variances 91
3.5.1 Confidence Intervals 91
3.5.2 Hypothesis Tests 92
3.5.3 Testing two means when variances are equal 93
3.6 Generalized Hypothesis Test Procedures 93
3.6.1 Steps in Hypothesis Testing 94
3.6.2 Examples of Hypothesis testing using the Wizard: 96
3.7 Exercises 99
Chapter 4 Analysis of Variance 100
4.1 One-Way ANOVA 100
4.2 The Randomized Complete Block Design 102
4.3 The Latin Square Design 103
4.4 Factorial Designs 104
4.5 Random and Mixed Models 106
4.6 2*K Factorial Designs 107
4.6.1 A Single Replication of 2*K Design 108
4.6.2 Fractional Factorial Designs 109
4.7 Nested Designs 112
4.8 Summary Table 113
4.9 ANOVA Wizard 115
4.10 – Exercises 117
Chapter 5 - Regression 118
5.1 Simple Regression 119
5.2 Multiple Regression 123
5.3 Indicator Variables 125
5.3.1 Indicator Variables with more than two Categories 125
5.3.2 Multiple Indicator Variables 126
5.4 Polynomial Regression 126
5.4.1 Quadratic Regression 127
5.4.2 Polynomial Models in Two or More Variables 127
5.5 Variance Stabilizing Transformations 128
5.6 Multiplicative Regression 130
5.6 Custom Designed Regression 131
5.7 Regression Summary 132
5.8 Regression Wizard 133
5.9 Exercises 136
Chapter 6 – Non-Parametric Statistics 137
6.1 Chi-Square Tests 137

6.1.1 Goodness of Fit 137
6.1.2 Independence 139
6.1.3 Chi-Square Test Wizard 140
6.2 Testing the median – Sign Test 144
6.2.1 Testing the median – Large Sample 144
6.2.2 Sign Test on Paired Data 145
6.3 The Wilcoxon Signed-Rank Test 146
6.4 The Mann-Whitney U Test 146
6.4.1 The Mann-Whitney U Test – Small Samples 146
6.4.2 The Mann-Whitney U Test – Large Samples 147
6.5 The Kruskal-Wallis Test – Three or more groups 148
6.6 Goodness-of-Fit Tests 148
6.6.1 The Kolmogorov Test 149
6.6.2 The Lillefors Test 151
6.6.3 The Smirnov Test 151
6.7 Exercises 152
Chapter 7 – Simulation 153
7.1 Apartment Rental Problem 153
7.2 Newsvendor Problem 153
7.3 Monte Hall Problem 154
7.4 Auto Insurance Problem 154
7.5 Piedmont Airlines Problem 156
7.6 Exercises 157
Chapter 8 – Decision Analysis 158
8.1 Exercises 159
Appendix A – Reference Card 160

Taming Statistics with TamStat

Chapter 0 – A Few Preliminaries

0.1 Installation

The desktop version of TamStat consists of a single executable file: tamstat.exe .   Load the file and TamStat will put up a session screen.

The HTML version of TamStat runs on both Windows and the Mac and Linux. Depending upon your operating system, follow the instructions below to install TamStat.

0.1.1 Windows Installation

- Download the file TamStat-Windows-Setup_vn.n.n.exe. Click on the downloaded file and follow the instructions.

0.1.2 Mac Installation

Welcome to TamStat Mac! With this document we want to provide some instructions for the (simple) installation. A more evolved installer for the Mac is currently in development, but at the moment you have to go through these steps manually.

1. unzip the file TamStat-Mac into a temporary folder. Looks like you've done that already, since you are reading this file which was contained in the ZIP-Archive. Well done!
2. install mac-64_18.0.xxxxx_unicode.pkg
3. move the TamStat-Folder into its future destination (your home-directory or the desktop) - no specific path required, as long as you read/write files in that path.
4. open the TamStat-folder, click on the file "tamstat" and press Command-I to open the "Get Info" dialog. In its bottom-section, activate the 3 cheeckboxes in the column "Execute".
   NB: depending on O/S Version, user permissions etc., this dialog may look different. But its essential that we changes those permissions! Alternatively, open the Terminal application. `cd ~/Desktop` will change into the desktop folder, then do another `cd` and use the name of the folder you unzipped TamStat to (probably "TamStat"). Before doing anything, let's check we're in the correct location: type the following command `ls -l | grep -v '^d'` - this shows a list of the files in that folder. You should see 3 entries, the last one being "tamstat". If this is true, let's change its attributes and make it executable: `chmod 755 tamstat`
   Done! (Apologies for the extra work and thanks for your patience! 😉)
5. from now on, simply navigate to the TamStat-folder and double-click "tamstat" to launch TamStat!

If you have any questions or suggestions, please feel free to share them with Stephen Mansour or Michael Baas (mbaas@dyalog.com)

0.1.3 Linux Installation

Download the file TamStat-Linux-vn.n.n.zip. Unzip the file and follow the instructions in readme.html.

0.2 APL Syntax

TamStat is written in Dyalog APL. The syntax is powerful and concise. A basic understanding of this syntax will go a long way in making you more productive in TamStat. In this section we will cover the basics of this syntax. For a more complete understanding of the APL language, we recommend you go to https://www.tryapl.org. If you are familiar with APL syntax, you can skip the remainder of this section.

Type in `2+3` in the session and TamStat will return the number `5`. It can't get much simpler than that.

```
      2+3
5
```

We can also assign the result of this operation to a variable name:

```
     A ← 2+3
     A
5
```

We can now use the variable `A` in an expression

```
     A+2
7
```

0.2.1 Data and Variables

Data have the following structures: a scalar (single value), a vector (list) or a matrix (table). We use the term variable to represent data which have been assigned a name. Variable names always begin with a capital letter. Here is an example of a scalar variable:

```
    S←3
```

A vector is a list of values:

```
    V←2 4 1
```

A matrix is a table of values:

```
    M←3 4 matrix 1 to 12
1  2  3  4
5  6  7  8
9 10 11 12
```

Variables can also be nested; that is a vector can consist of items which themselves are vectors:

```
    NV ← (2 3 4)(5 6)(7 8 9 10)
```

0.2.2 Functions

Functions take data as input and produce new data. Inputs to functions are known as arguments. Functions are denoted by familiar symbols or by names beginning with a lower-case letter. There are two basic types of functions: monadic and dyadic. A monadic function takes an argument on its right. An example of a monadic function is the square root function:

```
     sqrt 4
```

```
2
```

A dyadic function takes both a left argument and a right argument. Addition is dyadic function

```
     A ← 2 + 3
```

Let’s see what is going on:

```
    A         ←            2            +          3
    ^         ^            ^            ^          ^
    |         |            |            |          |
Result    Assignment      Left       Function     Right
                        Argument               Argument
```

Scalar functions operate on data in an item-by-item fashion. Thus:

```
     sqrt V
1.4142 2 1
     1 2 3 + 4 5 6
5 7 9
```

If one of the items is a scalar, it is applied to each item in the other argument:

```
     2 + 4 5 6
6 7 8
     4 5 6 + 10
14 15 16
```

Scalar dyadic functions require both arguments be the same length unless one of the arguments is a scalar:

```
      2 3 + 4 5 6
LENGTH ERROR
```

Scalar dyadic functions include the four arithmetic operations (`+`, `-`, `times` and `div`) as well as the power function:

```
      2 * 3
8
```

Not all functions are scalar functions. One example of a non-scalar function is catenate (`,`) which joins two vectors together:

```
       W←V,4 5
2 4 1 4 5
```

0.2.3 Order of operations

The left argument of a function is the variable immediately to its left. The right argument to a function is the result of the expression to its right. Thus:

```
       2 × 3 + 4
14
```

To override the order of operations, use parentheses:

```
       (2 × 3) + 4
10
```

If you are in doubt as to the order of operations, you may always add parentheses even if they are redundant. Thus, the following example will add 3 and 4 before multiplying even though the parentheses are redundant.

```
      2 × (3 + 4)
14
```

More complicated expressions such as $\frac{150-144}{10/\sqrt{25}}$ can be written as follows:

```
      (150-144)÷(10÷sqrt 25)
3
```

The last set of parentheses are redundant, but they clarify the expression.

0.2.4 Operators

Operators take functions as input and produce new functions. For example, the operator `each` applies a function to each item in a vector:

```
      sum 2 3 4
9
      sum each (2 3 4) (5 6 7)
9 18
```

We can apply the same operator to another function, catenate (,):

```
      2 ,each (4 5)(6 7 8) 9
2 4 5    2 6 7 8    2 9
```

The derivative in calculus is another example of an operator. In mathematics, we write $f'(x)$ (read f prime x). In TamStat we can write this as:

```
    f prime X
```

In TamStat, we calculate the natural log of 2:

```
      ln 2
0.6931471806
```

Now we apply the operator prime which calculates the derivative of the log function at 2:

$$g(x) = \ln x \qquad g'(x) = \tfrac{1}{x} \qquad g'(2) = \tfrac{1}{2}$$

```
      ln prime 2
0.5
```

We can apply the operator **`prime`** to another function **`sqrt`** to produce a different function:

$$h(x) = \sqrt{x} \qquad h'(x) = \frac{1}{2\sqrt{x}} \qquad h'(4) = \frac{1}{2\sqrt{4}} = \frac{1}{4}$$

```
      sqrt prime 4
0.25
```

0.2.5 Namespaces

A namespace is like a folder which contains variables, functions and/or operators instead of files. Namespaces, like variables begin with a capital letter. To find the names of the variables in a namespace, use the function "variables":

```
        variables NS
A B
```

To refer to any of the variables in a namespace, simply type the name of the namespace followed by a dot and the name of the variable:

```
        NS.A
2
        NS.B
5 2 3 1 4
        NS.A+NS.B
7 4 5 3 6
```

A namespace can also contain functions.

```
       NS.f ← {3 + 2 × ⍵}

       NS.f 2
7
```

The root namespace is #. In the desktop version of TamStat, you are in the root namespace, so you don't have to worry about this. If you are using the HTML version, you will automatically be inside the SD namespace. To refer to anything outside of this namespace, you must indicate the if a function or variable is in the root namespace you can refer to items outside the current namespace in the following way:

```
       #.SD         ⍝ Namespace SD
       #.SD.Height  ⍝ Variable Height in namespace SD
```

0.5 Exercises

1. A storage container is 30 feet long, 6 feet wide and 10 feet high. How many cubic feet of storage space are available?
2. You want 2 loaves of bread, 1 dozen eggs and 3 quarts of milk. Bread costs $3.25 a loaf, eggs cost $1.00 a dozen, and milk is $.75 a quart. How much money do you need?
3. A ten foot ladder is leaning against a house. The base of the ladder is 3 feet away from the house. How high does the ladder reach?
4. The future value of an investment of *P* dollars compounded quarterly is $P\left(1+\frac{r}{4}\right)^{4t}$ where r is the interest rate. If you invest $1500 compounded quarterly at 3.25%, how much money will you have in 7 years?
5. If you take out a $150,000, thirty-year fixed-rate mortgage at 4.25% what is your monthly payment? Use the formula:

   $$A = \frac{Br(1+r)^N}{(1+r)^N - 1}$$

   where *B* is the original balance, *r* is the monthly interest rate, 0.0425/12 and *N* is the term of the mortgage in months.

6. A major league pitcher can throw a 90-mile-per-hour fastball. If he is six feet tall and throws a ball directly upward, how high will it be after 1, 2, 3 and 4 seconds? Use the formula:

   $$h(t) = h_0 + v_0 t - 16t^2$$

   where h_0 is the height of the pitcher, v_0 is the initial velocity in feet per second and *t* is time in seconds.

Answers:

1.
```
ASCII:   30 * 6 * 10    or    product 30 60 10
APL:     30 × 6 × 10    or    ×/30 6 10
1800 cubic feet
```
2.
```
ASCII:   sum 2 1 3 * 3.25 1 0.75
APL:     2 1 3+.×3.25 1 0.75
$9.75
```
3.
```
ASCII:   sqrt (10^2)-3^2
APL:     (-/10 3*2)*0.5
9.539392014 feet
```
4.
```
ASCII:   1500*(1+0.0325/4)^4*7
APL:     1500×(1+0.0325÷4)*4×7
$1881.46
```
5.
```
ASCII:   R <- 0.0425/12
         N<- 12 * 30
         (150000*R*(1+R)^N)/((1+R)^N) - 1
APL:     R←0.0425÷12 ⋄ N←12×30
         (150000×R×(1+R)*N)÷¯1+(1+R)*N
$737.91
```
6.
```
ASCII:   T <- 1 2 3 4
         V0 <- 90 * 5280 / 3600
         6 + (V0 * T) - 32 * T^2
APL:     T←1 2 3 4 ⋄ V0←90×5280÷3600
         6 + (V0 × T) - 32 × T*2
106 142 114 22
```

Chapter 1 - Descriptive Statistics

Data can be organized in the following way:

A **record** is a set of values corresponding to a particular item. It corresponds to a row of the table.
A **variable** is a set of values corresponding to a particular attribute. It corresponds to a column of the table.
A **data set** is a collection of variables containing attributes for a set of entities. It corresponds to a table.
An **observation** is a single value. It corresponds to a cell in the table. See Figure 1 below.

KEY	Attribute 1	Attribute 2	Attribute 3	…	Attribute n
Item 1					
Item 2		Observation			
Item 3					
Item 4	←--------------------- Record ---------------------------→				
⋮					
Item m					

KEY	Attribute 1	Attribute 2	Attribute 3	…	Attribute n
Item 1			^		
Item 2			|		
Item 3			Variable		
⋮			|		
Item m			v		

Figure 1 - Observations, Records and Variables

In Table 2 on the next page, the items correspond to students. The variables are Party, Sex, Major, etc. A record consists of all the attributes of a particular student, whereas a variable contains one attribute for all students.

We can describe a data set by a **data dictionary** in Table 1 below which defines each variable explicitly:

Variable	Type	Scale of Measurement	Description
Student	Qualitative	Nominal	Number uniquely identifying student (Key Field)
Party	Qualitative	Nominal	D=Democrat, R=Republican, I=Independent
Sex	Qualitative	Nominal	F=Female, M=Male
State	Qualitative	Nominal	Home state; 2-digit postal code
Eyes	Qualitative	Nominal	Eye color
Height	Quantitative	Ratio	Height in inches
Weight	Quantitative	Ratio	Weight in pounds
ShoeSize	Quantitative	Interval	Shoe Size
Family	Quantitative	Ratio	Number of siblings
Hand	Qualitative	Nominal	L=Left-handed, R=Right-Handed, A=Ambidextrous
Car	Qualitative	Nominal	1=Student owns a car, 0=student doesn't own car
Pot	Qualitative	Ordinal	Answer to survey question: Do you think marijuana should be legalized? 1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly Agree

Table 1- Data Dictionary

Student	Party	Sex	State	Eyes	Height	Weight	ShoeSize	Family	Hand	Car	Pot
1	R	M	PA	Blue	72	220	11.5	3	R	1	4
2	D	F	PA	Brown	62	140	9	4	R	1	5
3	D	M	MD	Blue	69	195	11	0	R	0	4
4	D	M	PA	Blue	69	190	9.5	1	R	1	3
5	R	M	CT	Brown	70	150	10.5	1	L	1	5
6	R	M	NJ	Brown	66	125	8.25	2	R	1	3
7	D	M	NY	Brown	67	155	9	1	R	0	4
8	I	M	PA	Green	72	260	13	2	L	1	4
9	R	M	NY	Blue	72	155	10.5	2	R	1	4
10	D	M	CT	Brown	71	180	12	2	R	1	4
11	R	M	MD	Blue	71	160	11	1	R	1	2
12	R	M	PA	Blue	73	280	11.5	3	R	1	3
13	R	F	PA	Green	61	105	6.5	3	L	1	3
14	I	M	NY	Brown	65	120	7	1	R	0	5
15	R	F	NJ	Brown	69	150	9.5	1	R	1	4
16	R	F	PA	Blue	65	160	7.5	3	R	1	1
17	R	M	PA	Blue	73	185	12.5	1	R	1	2
18	I	M	NY	Brown	67	140	9	2	R	1	4
19	I	M	PA	Green	70	165	11.5	1	R	0	5
20	R	M	NJ	Other	68	210	12	1	R	1	3
21	I	F	NJ	Blue	55	142	8.5	3	R	1	4
22	I	M	NJ	Hazel	74	220	13	1	R	1	5
23	D	M	PA	Brown	71	165	11	2	L	0	3
24	I	M	CT	Blue	72	165	10	1	R	1	5
25	D	M	NY	Brown	75	205	13	2	L	1	5
26	R	M	NY	Blue	70	180	11.5	3	R	1	4
27	R	M	PA	Brown	71	139.5	11.5	1	R	1	1
28	R	M	PA	Brown	71	130	9.5	2	R	1	5
29	R	M	NJ	Hazel	71	160	10.5	3	R	1	4
30	I	M	MD	Blue	74	245	14	1	R	0	4
31	D	M	PA	Brown	61	100	6	2	R	1	3
32	R	F	PA	Brown	67	170	9.5	1	R	0	3
33	D	F	PA	Brown	64	150	8	2	R	1	2
34	I	F	NY	Brown	62	115	6	1	R	1	4
35	I	M	NY	Brown	75	195	12	2	R	0	4
36	D	F	NY	Brown	65.5	115	7	1	R	0	3
37	I	M	CT	Blue	72	185	11.5	0	R	1	3
38	D	M	PA	Hazel	71	225	11.5	1	R	1	3

Table 2: Student Data Set

Descriptive statistics involve summarizing data in one of three ways: Tables, summary statistics, and graphs. There are two broad categories of data: Qualitative and Quantitative. Qualitative or categorical data are usually (but not always) represented as text. Quantitative data must be represented as numeric.   We can represent qualitative data with a table.

1.1 Importing and Editing Databases in TamStat

The desktop and HTML versions can both use .csv files as source files for databases; however, the procedure for importing them into TamStat differ.   Both procedures are documented in this section.

1.1.2 Desktop Version

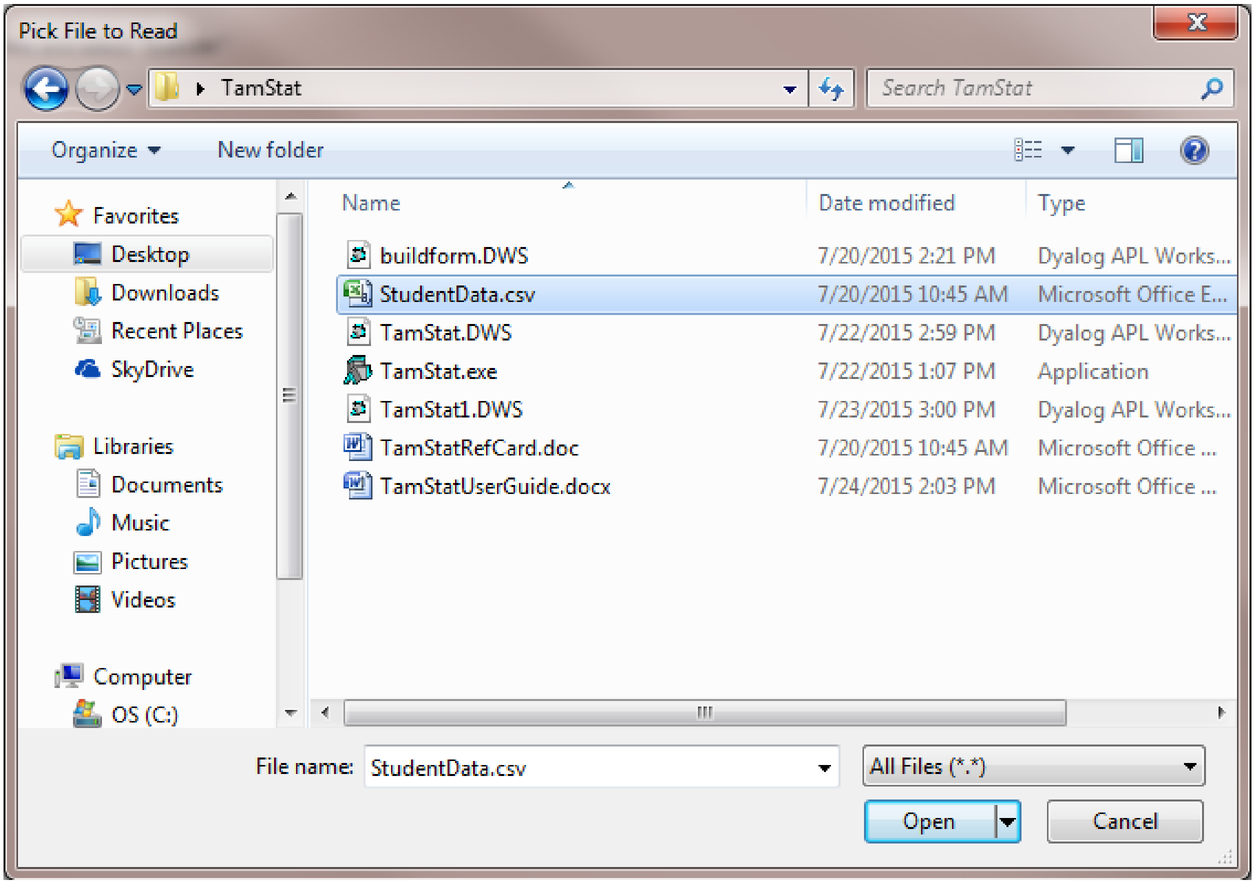
Let us import the data from the file `StudentData` which is a CSV (comma separated variable) file.   CSV files can be created in Excel. We can enter:

```
    SD← import '[pathname]/StudentData.csv'
```

If the pathname and file name are not specified as the right argument, the user can enter:

```
    SD← import ''
```

Then the user can navigate the directory and find the .csv file directly:



Name	Date modified	Type
buildform.DWS	7/20/2015 2:21 PM	Dyalog APL Works...
StudentData.csv	7/20/2015 10:45 AM	Microsoft Office E...
TamStat.DWS	7/22/2015 2:59 PM	Dyalog APL Works...
TamStat.exe	7/22/2015 1:07 PM	Application
TamStat1.DWS	7/23/2015 3:00 PM	Dyalog APL Works...
TamStatRefCard.doc	7/20/2015 10:45 AM	Microsoft Office ...
TamStatUserGuide.docx	7/24/2015 2:03 PM	Microsoft Office ...

File name: StudentData.csv

All Files (*.*)

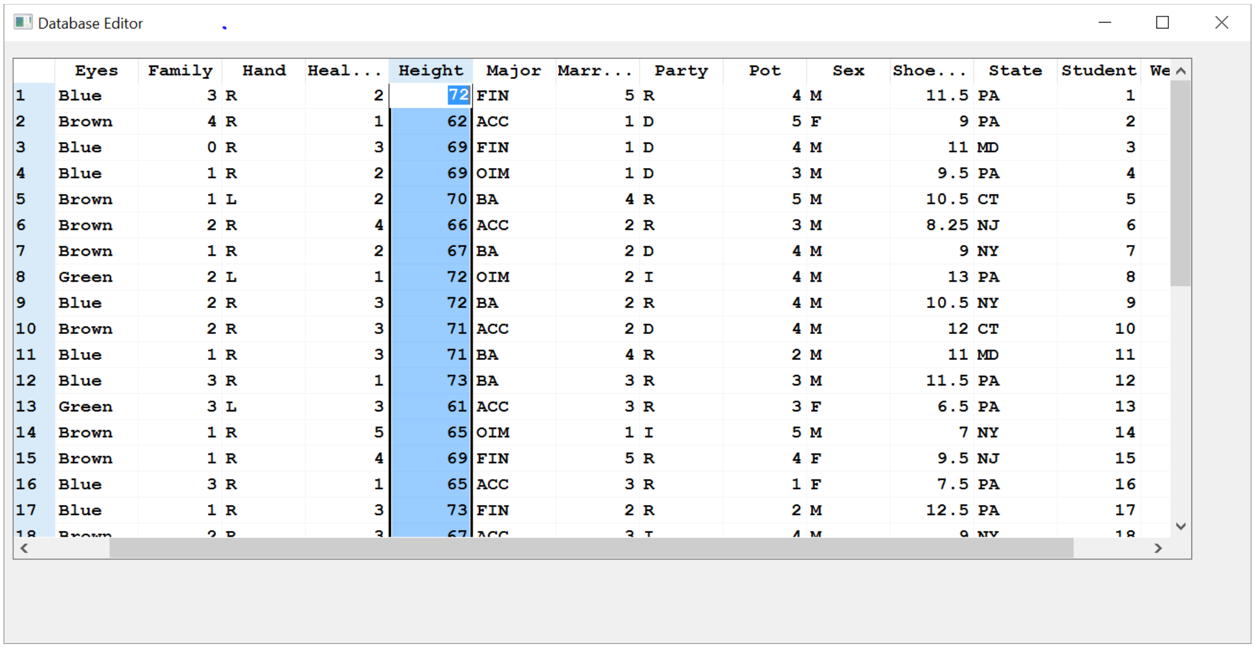
To obtain a list of variables in this data set, enter:

```
     variables SD
Car  Class  Eyes  Family  Hand  HealthCare  Height  Major Marriage  Party Pot
Religion  Sex  ShoeSize  State  Student  Weight
```

To edit an existing database, enter:

```
        SD ← editDatabase SD
```

The following screen will appear. You may change any value, then press x in the upper right hand corner to save.

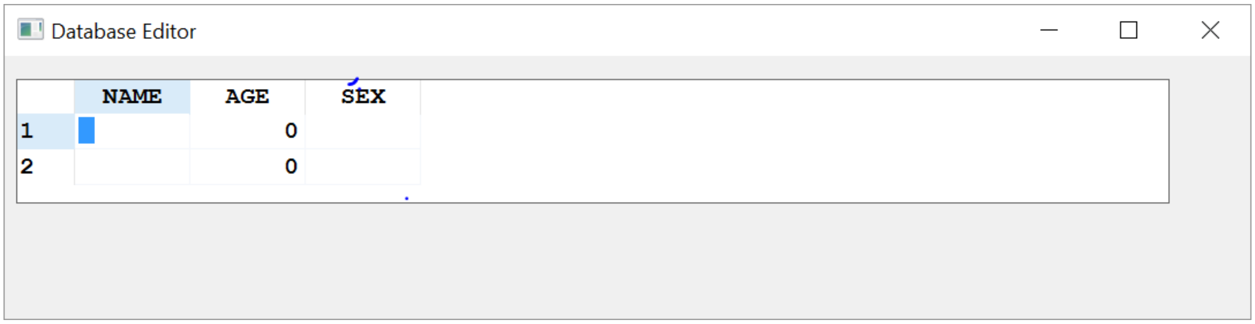


	Eyes	Family	Hand	Heal...	Height	Major	Marr...	Party	Pot	Sex	Shoe...	State	Student	We
1	Blue	3	R	2	72	FIN	5	R	4	M	11.5	PA	1	
2	Brown	4	R	1	62	ACC	1	D	5	F	9	PA	2	
3	Blue	0	R	3	69	FIN	1	D	4	M	11	MD	3	
4	Blue	1	R	2	69	OIM	1	D	3	M	9.5	PA	4	
5	Brown	1	L	2	70	BA	4	R	5	M	10.5	CT	5	
6	Brown	2	R	4	66	ACC	2	R	3	M	8.25	NJ	6	
7	Brown	1	R	2	67	BA	2	D	4	M	9	NY	7	
8	Green	2	L	1	72	OIM	2	I	4	M	13	PA	8	
9	Blue	2	R	3	72	BA	2	R	4	M	10.5	NY	9	
10	Brown	2	R	3	71	ACC	2	D	4	M	12	CT	10	
11	Blue	1	R	3	71	BA	4	R	2	M	11	MD	11	
12	Blue	3	R	1	73	BA	3	R	3	M	11.5	PA	12	
13	Green	3	L	3	61	ACC	3	R	3	F	6.5	PA	13	
14	Brown	1	R	5	65	OIM	1	I	5	M	7	NY	14	
15	Brown	1	R	4	69	FIN	5	R	4	F	9.5	NJ	15	
16	Blue	3	R	1	65	ACC	3	R	1	F	7.5	PA	16	
17	Blue	1	R	3	73	FIN	2	R	2	M	12.5	PA	17	
18	Brown	2	R	3	67	ACC	3	I	4	M	9	NY	18	

To create a new database from scratch, you need to list the column names. Indicate character fields by preceding the name with an ampersand:

```
      NEWDB←editDatabase'&NAME,AGE,&SEX'
```

Note that TAMSTAT creates exactly two rows. You can cursor down to create additional rows. Also note that the character fields are initialized with blanks and the numeric fields are initialized with zeroes. Note that the ampersand is not displayed for the character field names, because it is not actually part of the name.



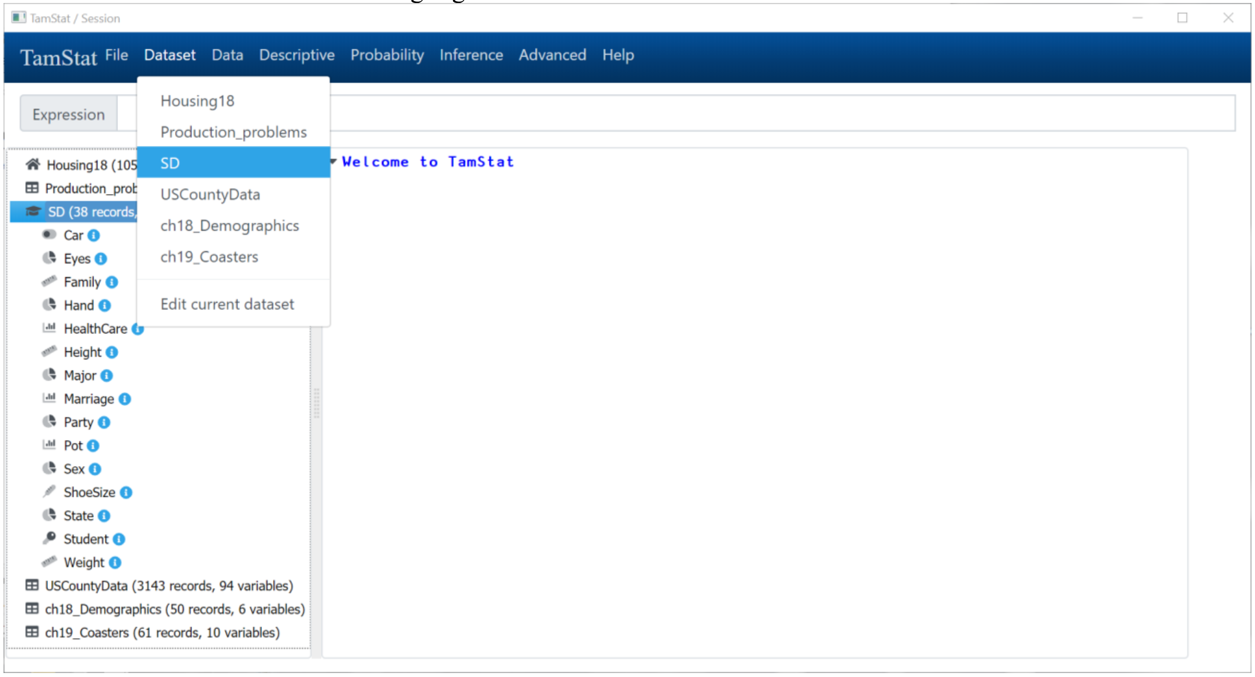
	NAME	AGE	SEX
1		0	
2		0	

1.1.2 HTML Version

In order to work with a database, you must first put a .csv file in the following folder:

```
[Path]\TamStat-Win-vX.X\TamStat-App\Data
```

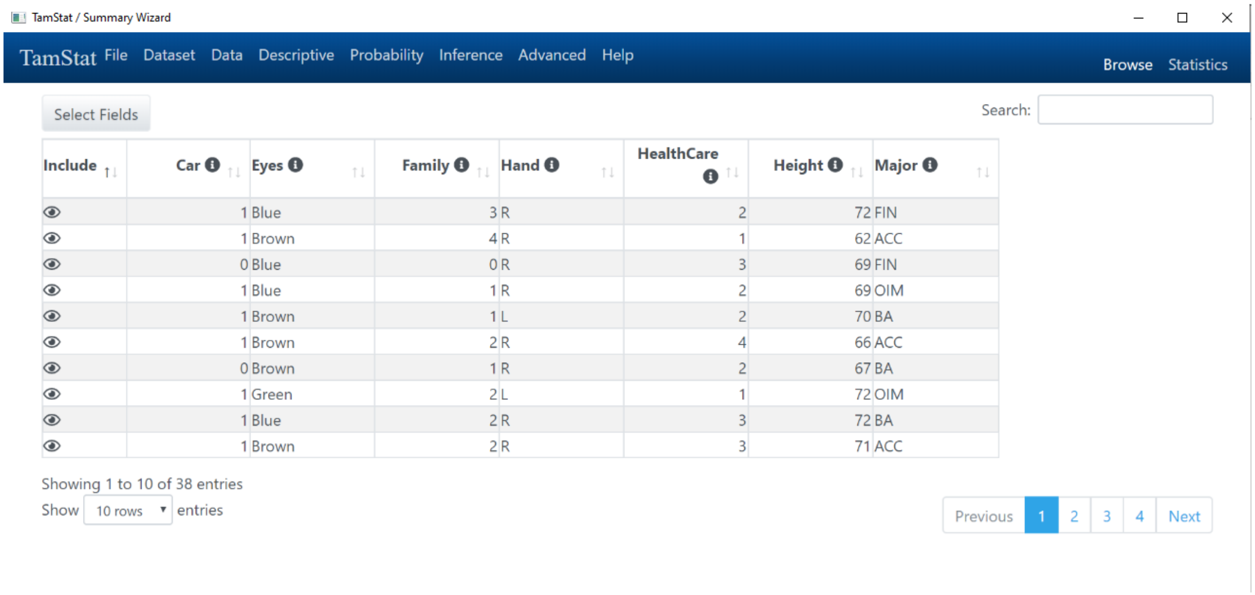
There are several sample databases shipped with TamStat as .csv files.   To work with a particular database, simply select “Dataset” from the menu and highlight the database of interest.



```
Welcome to TamStat
```

The database SD contains Student Data. Notice that when you select Data Session, the TreeView on the left lists all the databases in the Data folder; each database can be expanded to list all the variables within.

To view the database, just select `SummaryWizard` and the following screen will appear.   Use the `SelectFields` button to display the fields you are interested in:



Include	Car	Eyes	Family	Hand	HealthCare	Height	Major
👁	1	Blue	3	R	2	72	FIN
👁	1	Brown	4	R	1	62	ACC
👁	0	Blue	0	R	3	69	FIN
👁	1	Blue	1	R	2	69	OIM
👁	1	Brown	1	L	2	70	BA
👁	1	Brown	2	R	4	66	ACC
👁	0	Brown	1	R	2	67	BA
👁	1	Green	2	L	1	72	OIM
👁	1	Blue	2	R	3	72	BA
👁	1	Brown	2	R	3	71	ACC

Showing 1 to 10 of 38 entries

Show 10 rows entries

1.1 Tables

1.1.1 Categorical Data – frequency, relative frequency

Let’s look at the qualitative variable STATE. We could simply display the variable as a list:

```
      #.SD.State

 PA  PA  MD  PA  CT  NJ  NY  PA  NY  CT  MD  PA  PA  NY  NJ  PA  PA  NY  PA
NJ  NJ  NJ  PA  CT  NY  NY  PA  PA  NJ  MD  PA  PA  PA  NY  NY  NY  CT  PA
```

But it would be more useful to create a frequency distribution and count the number in each category:

```
      frequency #.SD.State

 CT   4
 MD   3
 NJ   6
 NY   9
 PA  16
```

To obtain the relative frequency, we simply divide the frequencies by the total number of items in the sample:

```
      relative frequency #.SD.State

 CT  0.10526
 MD  0.078947
 NJ  0.15789
 NY  0.23684
 PA  0.42105
```

Or if you prefer to express everything in percentages:

```
      percent frequency #.SD.State
 CT  10.526
 MD   7.8947
 NJ  15.789
 NY  23.684
 PA  42.105
```

Since nominal data are not ordered, the frequency function automatically sorts them alphabetically.   We can also order them by frequency by supplying a left argument of -1. This is known as a Pareto analysis:

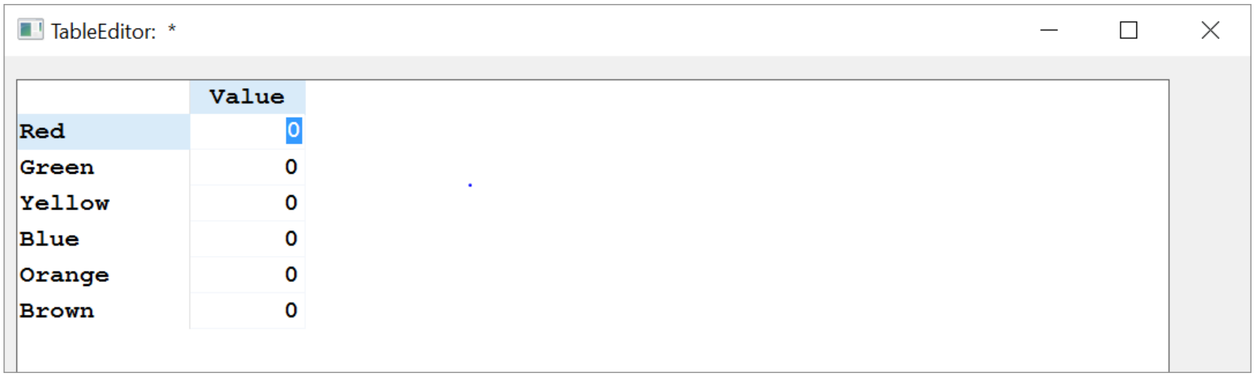
```
      ¯1 frequency #.SD.State

 PA  16
 NY   9
 NJ   6
 CT   4
 MD   3
```

To obtain a full report, we apply the report function:

```
       report ¯1 frequency #.SD.State

  Category Frequency  Cum Freq   Percent    Cum Pct
 --------- ---------  ---------  ---------  ---------
 PA               16         16     42.11%     42.11%
 NY                9         25     23.68%     65.79%
 NJ                6         31     15.79%     81.58%
 CT                4         35     10.53%     92.11%
 MD                3         38      7.89%    100.00%
 --------- ---------  ---------  ---------  ---------
   Total          38         38    100.00%    100.00%
```

If you do not have access to the raw data and want to create a frequency table, use the `editTable` function. This is only available in the desktop version:

```
       FT←editTable 'Red,Green,Yellow,Blue,Orange,Brown'
```



TableEditor: *

	Value
Red	0
Green	0
Yellow	0
Blue	0
Orange	0
Brown	0

1.1.2 Contingency tables

A contingency table is a summary of two categorical variables. The rows represent values of the first variable while the columns represent values of the second variable. Let us create a very simple contingency table for the two variables Sex and Party using `the frequency` function with two variables:

```
        frequency #.SD.Sex #.SD.Party
    D  I   R
 F  3  2   4
 M  8  9  12
```

The rows of the table represent the variable Sex, showing Female (F) and Male (M); while the columns show the political affiliations: Democrat (D), Independent (I) and republican (R). The values in each cell represent the number in each category. Thus there are 3 students who are both Female (F) and Democrat (D).
Now it is often useful to include the totals for each row and column as well as the grand total. To accomplish this we can use the `show` operator (with a left argument of 0) to display this:

```
       frequency show #.SD.Sex #.SD.Party

 Count      |         D         I         R |    Total
 -------------------------------------------|--------
 F          |         3         2         4 |        9
 M          |         8         9        12 |       29
 -------------------------------------------|--------
 Total      |        11        11        16 |       38
```

Notice the row and column totals. These are known as marginal frequencies. If you look at just the first and last columns you will see a frequency distribution for Sex. If you look at just the first and last rows you will see the frequency distribution for Party. For comparison, we show the frequency function applied to the two variables Sex and Party separately:

```
     frequency #.SD.Sex
F   9
M  29

      frequency #.SD.Party
D  11
I  11
R  16
```

We can generate a relative or percent frequency distribution by dividing by the grand total of 38. To do this in TamStat, we must supply a left argument of 3:

```
       3 frequency show #.SD.Sex #.SD.Party

Total %   |          D         I         R |   Total
-------------------------------------------|--------
F         |      7.89%     5.26%    10.53% |  23.68%
M         |     21.05%    23.68%    31.58% |  76.32%
-------------------------------------------|--------
Total     |     28.95%    28.95%    42.11% | 100.00%
```

Now the first and last columns produce a relative or percent frequency distribution for Sex. Similarly, the first and last rows produce a relative or percent frequency distribution for Party. Again, we can apply the frequency function (with the show operator) to each variable separately to show this. Look at the Category and Percent columns for each table below and compare to the totals above.

```
       frequency show #.SD.Sex

 Category Frequency  Cum Freq   Percent   Cum Pct
--------- --------- --------- --------- ---------
F                 9         9    23.68%    23.68%
M                29        38    76.32%   100.00%
--------- --------- --------- --------- ---------
  Total                    38   100.00%   100.00%

      frequency show #.SD.Party

 Category Frequency  Cum Freq   Percent   Cum Pct
--------- --------- --------- --------- ---------
D                11        11    28.95%    28.95%
I                11        22    28.95%    57.89%
R                16        38    42.11%   100.00%
--------- --------- --------- --------- ---------
  Total                    38   100.00%   100.00%
```

To compare differences between male and female students, we can generate row percentages using a left argument of 1. Notice that the row totals are all 100%

```
      1 frequency show #.SD.Sex #.SD.Party
```

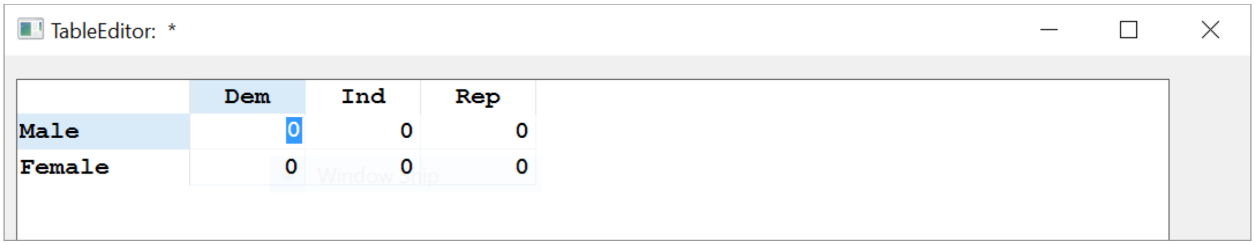
```
Row %      |          D          I          R |   Total
-----------------------------------------------|--------
F          |     33.33%     22.22%     44.44% | 100.00%
M          |     27.59%     31.03%     41.38% | 100.00%
-----------------------------------------------|--------
Total      |     28.95%     28.95%     42.11% | 100.00%
```

To compare differences between Parties, we can generate column percentages with a left argument of 2. Notice that all the column totals are 100%

```
      2 frequency show #.SD.Sex #.SD.Party

Column %   |          D          I          R |   Total
-----------------------------------------------|--------
F          |     27.27%     18.18%     25.00% |  23.68%
M          |     72.73%     81.82%     75.00% |  76.32%
-----------------------------------------------|--------
Total      |    100.00%    100.00%    100.00% | 100.00%
```

If you do not have access to the raw data and want to create a contingency table, you can also use the `editTable` function. (This is only available in the desktop version).

```
      CT←editTable 'Male,Female' 'Dem,Ind,Rep'
```



TableEditor: *

	Dem	Ind	Rep
Male	0	0	0
Female	0	0	0

1.1.3 Numeric Data – frequency, cumulative frequency, binning

Quantitative data can be handled in a similar way. The first column is the value and the second column is the frequency

```
    frequency #.SD.Family
0  2
1 17
2 11
3  7
4  1
```

For ordinal data and above, we can also represent data by a cumulative frequency:

```
      cumulative frequency #.SD.Family
0  2
1 19
2 30
3 37
4 38
```

Using the `show` operator, we can create a table showing relative and cumulative frequencies:

```
     frequency show #.SD.Family
  Category Frequency  Cum Freq   Percent   Cum Pct
 --------- --------- --------- --------- ---------
     0             2         2     5.26%     5.26%
     1            17        19    44.74%    50.00%
     2            11        30    28.95%    78.95%
     3             7        37    18.42%    97.37%
     4             1        38     2.63%   100.00%
 --------- --------- --------- --------- ---------
   Total          38        38   100.00%   100.00%
```

Sometimes there are too many unique values for a frequency table to be practical. In that case the frequency function will automatically group the data and display the midpoint of each range:

```
  frequency #.SD.Height
54  1
60  2
63  3
66  7
69  7
72 14
75  4
```

To display a full report showing frequencies, relative frequencies, cumulative frequencies and totals, use the report function:

```
    report frequency #.SD.Height

      From        To Frequency  Cum Freq   Percent   Cum Pct
 --------- --------- --------- --------- --------- ---------
    51.000 -  56.999         1         1     2.63%     2.63%
    57.000 -  62.999         2         3     5.26%     7.89%
    60.000 -  65.999         3         6     7.89%    15.79%
    63.000 -  68.999         7        13    18.42%    34.21%
    66.000 -  71.999         7        20    18.42%    52.63%
    69.000 -  74.999        14        34    36.84%    89.47%
    72.000 -  77.999         4        38    10.53%   100.00%
 --------- --------- --------- --------- --------- ---------
   Total                    38         1   100.00%   100.00%
```

If you prefer to change the cell widths, you can supply a left argument:

```
    5 frequency #.SD.Height
55  1
60  5
65 11
70 19
75  2
```

1.2 Summary Statistics

Can you describe your summer vacation in a single word? Some may say “relaxing”, “awesome” or “hot”. A summary function is similar to this. It allows us to describe a variable or set of data with a single value. The basic mathematical form of a summary function is:

$$y = f(x_1, x_2, \cdots, x_n)$$

We can also think of a summary function as a mapping from a vector to a scalar. Thus:

$$\vec{\mathbf{x}} = (x_1, x_2, \cdots, x_n) \quad y = f(\vec{\mathbf{x}})$$

A parameter is a summary function applied to a population. A statistic is a summary function applied to a sample. Some statistics measure the center of the data, while others measure the spread or dispersion. There are also measures of position and measures of shape.

1.2.1 Measures of Quantity

One simple way to describe a data set or variable is to count the number of items in it. This is known as the population or sample size and is indicated by *N* or *n* respectively. In TamStat we can find the sample size by applying the count function to any variable in the database:

```
     count #.SD.Height
38
```

Another way is to add up all the values:

$$\mathrm{y} = f(\vec{\mathbf{x}}) = \sum_{i=1}^{n} x_i$$

We can do this in TamStat by using the `sum` function:

```
     sum #.SD.Height
2613.5
```

Another useful summary function is the sum-of-squares:

$$\mathrm{y} = f(\vec{\mathbf{x}}) = \sum_{i=1}^{n} x_i^2$$

which can be obtained by using:

```
    sumSquares #.SD.Height
180488.25
```

1.2.1 Measures of Center

There are three ways to measure the center of the data. The first is the mean or average, which is the sum divided by the count:

```
      mean #.SD.Height
68.776
```

The middle value after the data are sorted is the median:

```
      median #.SD.Height
70
```

To see how the median was obtained, we use the `show` operator:

```
      median show #.SD.Height
55 61 61 62 62 64 65 65 65.5 66 67 67 67 68 69 69 69 70 70 | 70 71 71 71 71 71
71 71 72 72 72 72 72 73 73 74 74 75 75
```

And the most frequently-occurring value is the mode.

```
      mode #.SD.Height
71
```

Note that the mode also applies to categorical data; most students are from Pennsylvania:

```
      mode #.SD.State
PA
```

The mean is much more sensitive to extreme values than the median. Suppose someone enters a height of 680 for a student instead of 68 by mistake. The mean height which was 68.776 now balloons to:

```
      mean #.SD.Height,680
84.44871795
```

However, if we apply the median to a data set with outliers, it remains relatively unchanged:

```
      median #.SD.Height,680
70
```

In some cases it may be slightly but not significantly higher:

```
      median #.SD.Height,680 680
70.5
```

1.2.2 Measures of spread

Two data sets can have the same mean but can be completely different. For example the two variables below are different, but they have the same mean:

```
     A←100 100 100
     B←0 100 200
     mean A
100
     mean B
100
```

Perhaps we can get a better measure of spread if we subtract the smallest value from the largest value. This is known as the range:

```
     (range A)(range B)
0 200
```

But the range is very sensitive to outliers. We can measure the average difference between each value and the mean; unfortunately this value is always zero because the plusses cancel out the minuses:

```
      mean (A-mean A)
0
      mean (B-mean B)
0
```

We could take the average *distance* (absolute value of the difference) between each value and the mean. This would give us a more reasonable measure because distances are always positive:

```
      mean |A-mean A
0
      mean |B-mean B
66.667
```

This is better, but the absolute value is difficult to work with mathematically. Another way to ensure positive values is to square the differences. This is known as the variance.

```
      mean (A-mean A)*2
0
      mean (B-mean B)*2
6666.7
```

The problem with the variance is that the result is in squared units; to express the spread in the same units as the mean, we take the square root of the variance, better known as the standard deviation:

```
      sqrt mean (B-mean B)*2
81.65
```

Let's apply the variance and standard deviation to some of our sample data:

```
     var #.SD.Height
20.036
     sdev #.SD.Height
4.4762
```

To see how the variance and standard deviation are calculated, use the `show` operator:

```
     var 2 8 5 2 3
6.5
     sdev 2 8 5 2 3
2.5495
     var show 2 8 5 2 3
                _      _
     n   x   x-x   (x-x)²
--------------------------
     1   2    ¯2    4
     2   8     4   16
     3   5     1    1
     4   2    ¯2    4
     5   3    ¯1    1
--------------------------
 Total  20     0   26
  Mean   4   Var    6.5
            Sdev    2.5495
```

Believe it or not, the variance can be defined for qualitative data. According to Kader and Perry, the variance is defined as: $V = 1 - \sum_{i=1}^{k} n_i^2 / \left(\sum_{i=1}^{k} n_i\right)^2$ where n_i is the number of occurrences of category *i* in the sample. In TamStat we can simply apply the var function to a qualitative variable:

```
     var #.SD.State
0.7243767313
```

1.2.3 Measures of position

The median is also a measure of position. It is the value which is greater than half of all values. The median of the lower half of the data is known as the first quartile; the median itself is known as the second quartile, and the median of the upper half of the data is known as the third quartile. Let's find the quartiles of the height data:

```
      1 quartile #.SD.Height
66
      3 quartile #.SD.Height
72
      quartile show #.SD.Height
  55 61 61 62 62 64 65 65 65.5  ( 66 )  67 67 67 68 69 69 69 70 70  |  70 71
71 71 71 71 71 71 72  ( 72 )  72 72 72 73 73
```

The difference between the third quartile and the first quartile is known as the inter-quartile range or IQR. This is a better measure of spread than the range because it is much less sensitive to outliers.

```
      iqr #.SD.Height
6
```

While the range and variance are both extremely sensitive to outliers, `iqr` remains relatively stable:

```
      range #.SD.Height
20
      range #.SD.Height,680
625
      var #.SD.Height
20.03645092
      var #.SD.Height,680
9598.852564
      iqr #.SD.Height,680
6
```

We can also divide a data set into 100 percentiles. The kth percentile is the value which exceeds k% of all values:

```
       20 percentile #.SD.Height
65
       95 percentile #.SD.Height
75
```

The inverse of the percentile function is percentile rank. Given the value of the data, find the percentile:

```
       65 75  percentileRank #.SD.Height
18 97
```

1.2.4 Measures of shape

Some data are symmetric, while others may be skewed right or left.

```
       skewness #.SD.Height
¯1.0497
       skewness #.SD.Family
0.4816
```

The number of siblings has a positive skew, meaning there is a right tail. There is a lower limit to the number of siblings (0), but no upper limit.

1.2.5 Measures of Association

Some statistics involve more than one variable. When two variables are involved, these are of the form:

$$\vec{\mathbf{x}} = (x_1, x_2, \cdots, x_n) \quad \vec{\mathbf{y}} = (y, y_2, \cdots, y_n) \quad z = f(\vec{\mathbf{x}}, \vec{\mathbf{y}})$$

We can measure the association between two variables by calculating the covariance. The covariance is a generalization of the variance: $s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$

What is the covariance between height and weight?

```
     #.SD.Height cov #.SD.Weight
124.23
```

The covariance of a variable with itself is simply the variance:

```
     #.SD.Height cov #.SD.Height
20.036
     var #.SD.Height
20.036
```

We can normalize the covariance by dividing it by the product of the standard deviations of the two variables:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum(x-\bar{x})\,(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Correlation has the advantage of being unitless; that is two variables that are in a perfect upward-sloping straight line have a correlation of 1; two variables that form a downward-sloping straight line have a correlation of -1, and two variables that have no relation to each other have a correlation of 0. Two variables that have a strong linear relationship have a correlation close to 1 or -1; two variables that have a very weak linear relationship have a correlation close to 0.

```
      #.SD.Height corr #.SD.Weight
0.65807
```

This shows that there is a moderate correlation between height and weight. The correlation between height and shoe size is stronger:

```
      #.SD.Height corr #.SD.ShoeSize
0.82689
```

As might be expected, there is little or no correlation between weight and family size, as it is very close to 0:

```
      #.SD.Weight corr #.SD.Family
¯0.051639
```

The correlation of a variable with itself is 1:

```
      #.SD.Height corr #.SD.Height
1
```

Adding a constant or multiplying by a positive constant has no effect on the correlation:

```
      #.SD.Height corr #.SD.Weight×2
0.65807
      #.SD.Height corr #.SD.Weight+3
0.65807
```

Multiplying by a negative constant negates the correlation:

```
      #.SD.Height corr #.SD.Weight×-3
¯0.65807
```

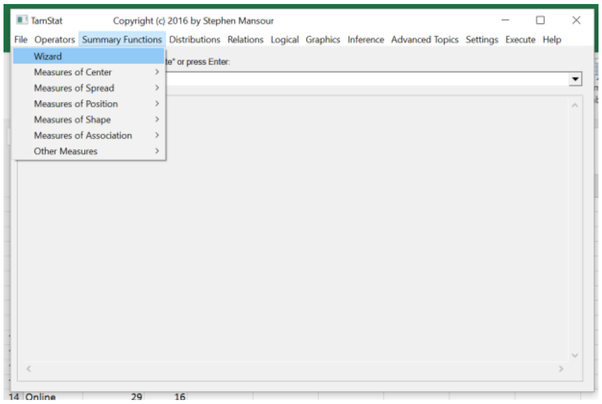
Using small data sets and the show operator illustrates how covariance and correlation are calculated:

```
     2 8 5 2 3 cov 1 7 3 4 1
5
     2 8 5 2 3 corr 1 7 3 4 1
0.78762
     2 8 5 2 3 cov show 1 7 3 4 1

     n   x    y  x-x̄   y-ȳ  (x-x̄)(y-ȳ)   (x-x̄)²  (y-ȳ)²
--------------------------------------------------------
     1   2  1    ¯2  ¯2.2     4.4       4         4.84
     2   8  7      4   3.8    15.2      16        14.44
     3   5  3      1  ¯0.2    ¯0.2       1         0.04
     4   2  4     ¯2   0.8    ¯1.6       4         0.64
     5   3  1     ¯1  ¯2.2     2.2       1         4.84
--------------------------------------------------------
 Total  20 16      0   0      20        26        24.8
  Mean   4  3.2    0   Cov     5         6.5       6.2
                      Corr     0.78762   2.5495    2.49
```

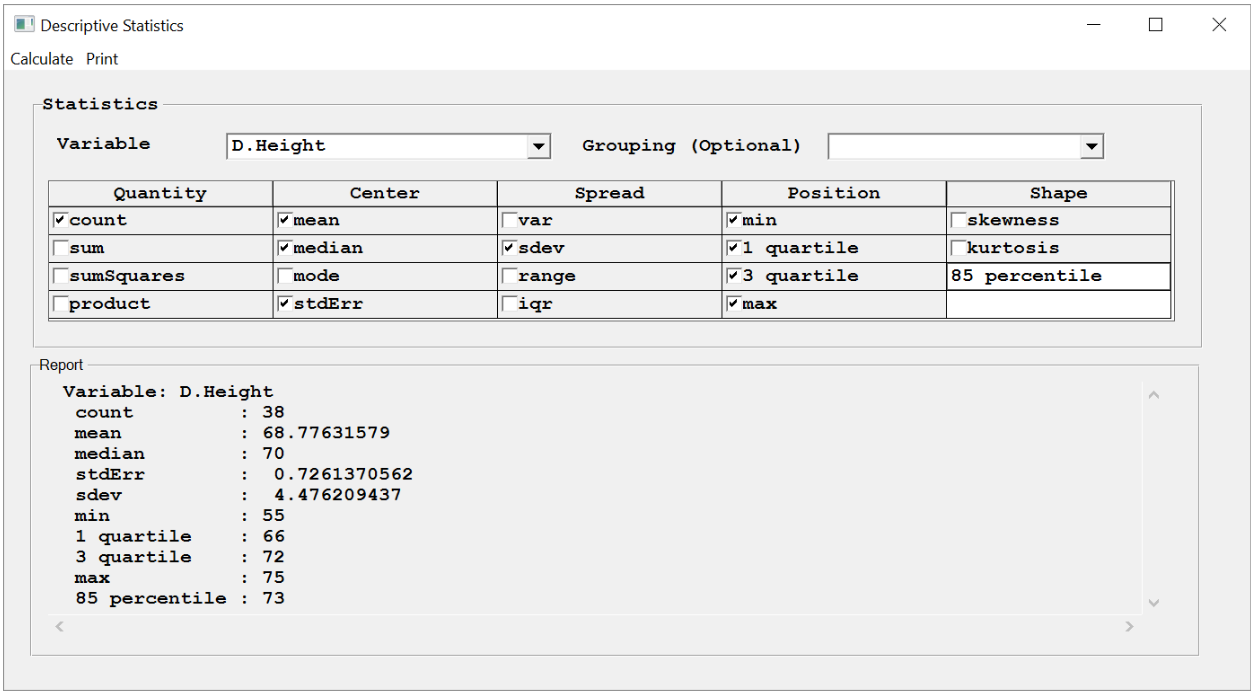
1.2.6 Summary Function Wizard

1.2.6.1 Desktop Version

To obtain summary statistics for a variable, one can use the Summary Function Wizard. From the main menu, select “Summary Functions”, then select “Wizard”.



The following screen will appear. The summary functions are organized by measures of quantity, center, spread, position and shape. Two blank cells in the lower right hand corner of the input box allow you to enter summary functions not listed. For example you could enter “`85 percentile`”, to obtain the 85th percentile of heights:

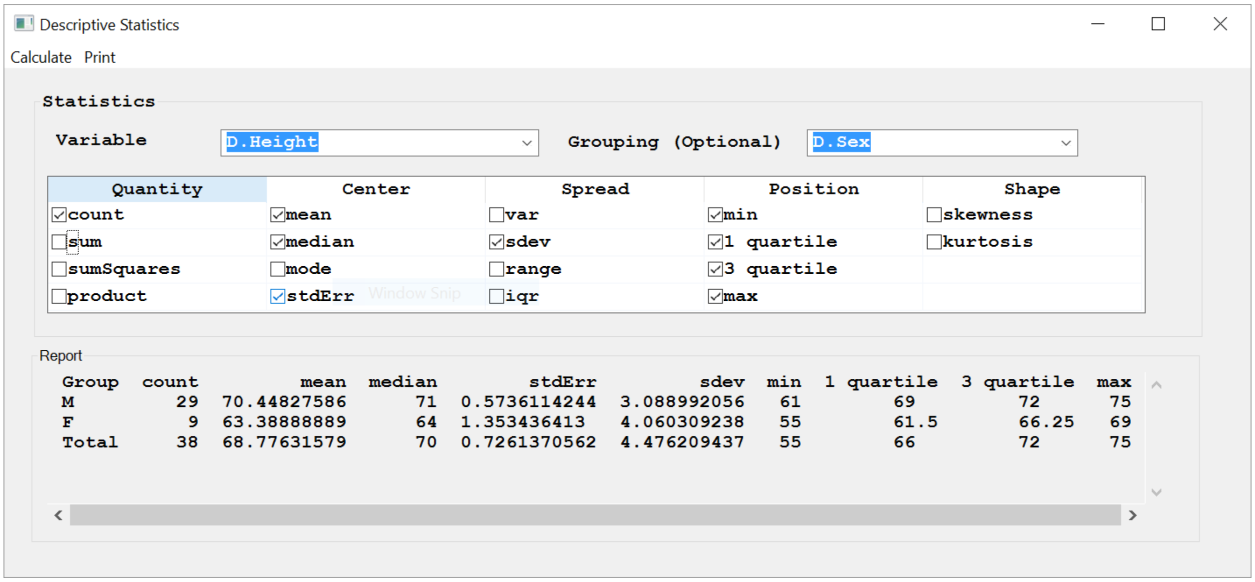


Variable: D.Height   Grouping (Optional):

Quantity	Center	Spread	Position	Shape
☑ count	☑ mean	□ var	☑ min	□ skewness
□ sum	☑ median	☑ sdev	☑ 1 quartile	□ kurtosis
□ sumSquares	□ mode	□ range	☑ 3 quartile	85 percentile
□ product	☑ stdErr	□ iqr	☑ max	

```
Variable: D.Height
 count         : 38
 mean          : 68.77631579
 median        : 70
 stdErr        :  0.7261370562
 sdev          :  4.476209437
 min           : 55
 1 quartile    : 66
 3 quartile    : 72
 max           : 75
 85 percentile : 73
```

To obtain summary statistics for various groups within a dataset, select the appropriate numeric variable along with a grouping variable. Grouping variables should be qualitative. The following example shows how to separate out height statistics between male and female students. Note the totals are also included.



Statistics

Variable: D.Height

Grouping (Optional): D.Sex

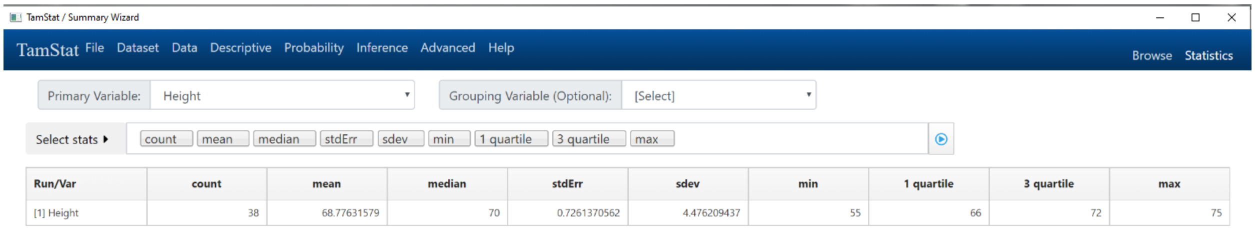
Quantity	Center	Spread	Position	Shape
☑count	☑mean	□var	☑min	□skewness
□sum	☑median	☑sdev	☑1 quartile	□kurtosis
□sumSquares	□mode	□range	☑3 quartile	
□product	☑stdErr	□iqr	☑max	

Report

```
Group   count         mean  median        stdErr         sdev  min  1 quartile  3 quartile  max
M          29  70.44827586      71  0.5736114244  3.088992056   61          69          72   75
F           9  63.38888889      64  1.353436413   4.060309238   55        61.5       66.25   69
Total      38  68.77631579      70  0.7261370562  4.476209437   55          66          72   75
```

1.2.6.2 HTML Version

Select Descriptive, then Summary Wizard, then at the far upper right, select Statistics.
Select the Primary Variable, e.g. "Height" then select the small blue circled arrow to display the results:


TamStat / Summary Wizard
TamStat File Dataset Data Descriptive Probability Inference Advanced Help
Browse Statistics

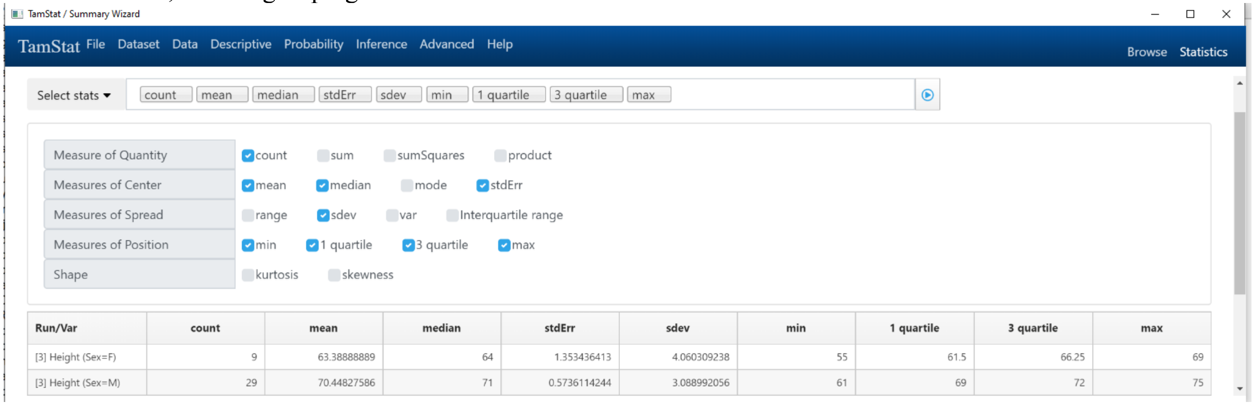
Primary Variable: Height

Grouping Variable (Optional): [Select]

Select stats ▸ count mean median stdErr sdev min 1 quartile 3 quartile max

Run/Var	count	mean	median	stdErr	sdev	min	1 quartile	3 quartile	max
[1] Height	38	68.77631579	70	0.7261370562	4.476209437	55	66	72	75

For more detail, use the grouping variable and the Select Stats button:


TamStat / Summary Wizard
TamStat File Dataset Data Descriptive Probability Inference Advanced Help
Browse Statistics

Select stats ▾ count mean median stdErr sdev min 1 quartile 3 quartile max

Measure of Quantity	☑count	□sum	□sumSquares	□product
Measures of Center	☑mean	☑median	□mode	☑stdErr
Measures of Spread	□range	☑sdev	□var	□Interquartile range
Measures of Position	☑min	☑1 quartile	☑3 quartile	☑max
Shape	□kurtosis	□skewness		

Run/Var	count	mean	median	stdErr	sdev	min	1 quartile	3 quartile	max
[3] Height (Sex=F)	9	63.38888889	64	1.353436413	4.060309238	55	61.5	66.25	69
[3] Height (Sex=M)	29	70.44827586	71	0.5736114244	3.088992056	61	69	72	75

1.3 Graphics

1.3.1 Pie Charts

A simple way to display qualitative data for a single variable is with a pie chart:

```
     G←pieChart #.SD.State
     G.Output

               * * * * * *                     Legend
           *.*......|      * *                 ┌─────────┐
        **..........|          **              │  CT    4│
     **.............|         %%%%**           │% MD    3│
    *...............|        %%%%%%%*          │: NJ    6│
   *................|      %%%%%%%%%%*         │@ NY    9│
  *.................|     %%%%%%%%%%%:*        │. PA   16│
 *..................|   %%%%%%%%:::::::*       └─────────┘
*...................| %%%%%:::::::::::::*
*...................|%%:::::::::::::::::*
*...................|@::::::::::::::::::*
*...................@@@@@:::::::::::::::*
*..................@@@@@@@@@::::::::::::*
 *................@@@@@@@@@@@@@::::::::*
  *..............@@@@@@@@@@@@@@@@@::::*
   *............@@@@@@@@@@@@@@@@@@@@@*
    *..........@@@@@@@@@@@@@@@@@@@@@*
     **.......@@@@@@@@@@@@@@@@@@@@**
        **..@@@@@@@@@@@@@@@@@@@**
           *@*@@@@@@@@@@@@@*@*
               * * *@* * *
```

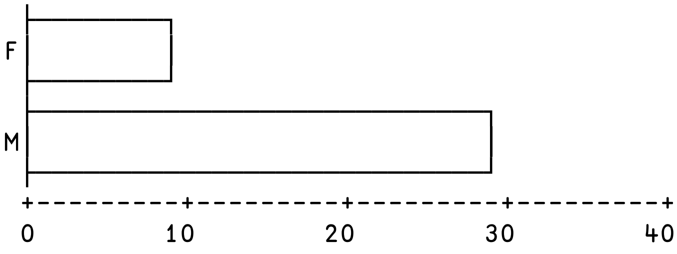
1.3.2 Bar Charts

Another way to represent qualitative data is with a bar chart. The `barChart` function can take either a variable or frequency table as input. Here we first create a frequency table:

```
      TABLE←frequency #.SD.Sex
 F   9
 M  29
```

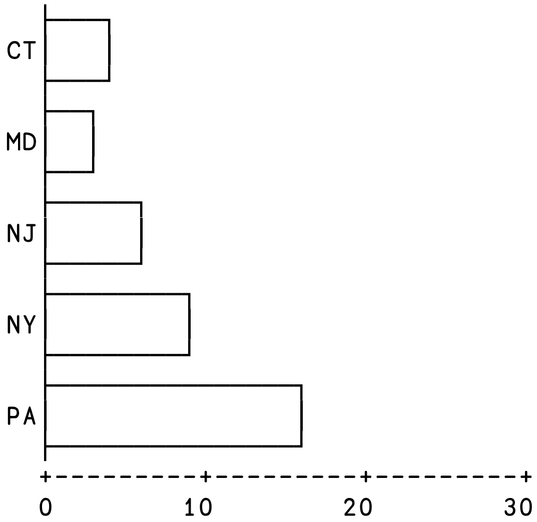
Then we create a bar chart from the table. Note we can use the `show` operator to display the chart immediately.:

```
      barChart show TABLE
```



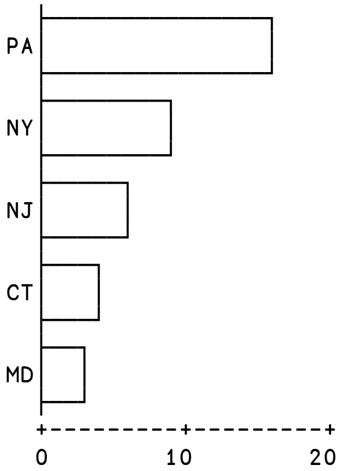
The barChart function can also take a categorical variable directly as input:

```
barChart show #.SD.State
```



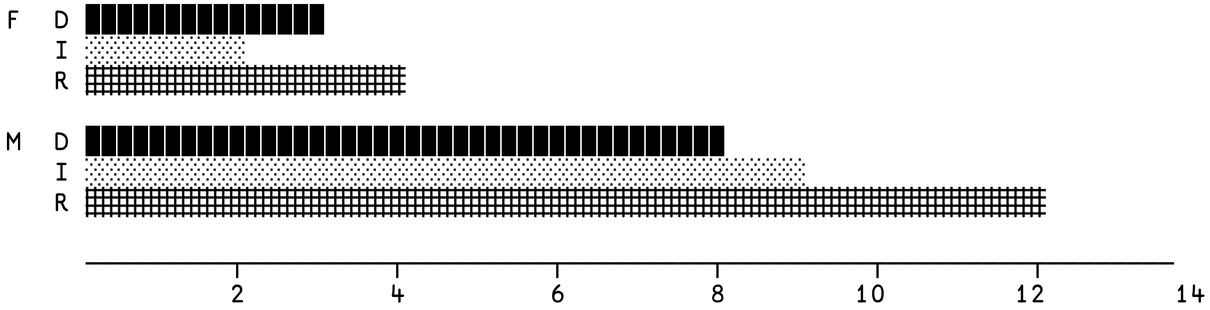
To create a Pareto chart (sorting by frequency), first create a frequency table:

```
barChart ¯1 frequency show #.SD.State
```



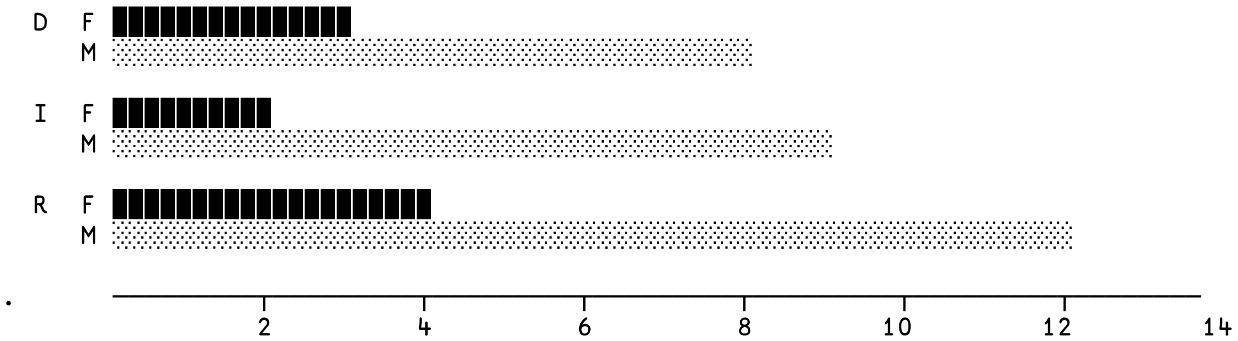
Bar charts can also be used to display two variables. This is known as a cluster bar chart:

```
barChart show #.SD.Sex #.SD.Party
```



The order of the two variables is important. The first variable is the major category, while the second variable iw clustered within it. We get a different chart when we reverse the order of Sex and Party:

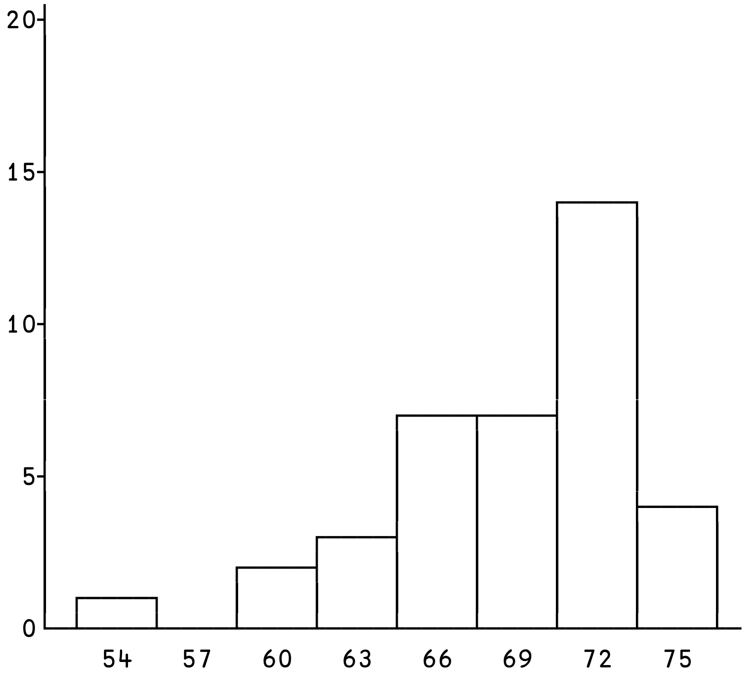
```
barChart show #.SD.Party #.SD.Sex
```



1.3.3 Histograms

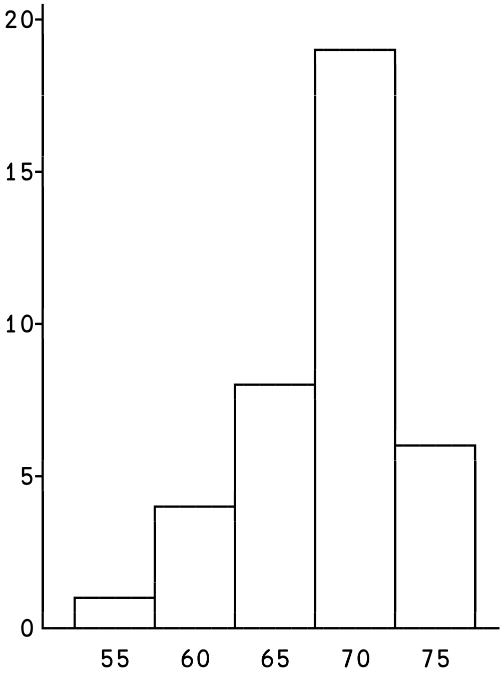
For continuous variables, the histogram is one of the most important graphical representations. It differs from a bar chart in that the bars are contiguous to reflect the continuity of the data. The value displayed below each bar is the midpoint of the data. To create a histogram, use the histogram function:

```
G←histogram #.SD.Height
G.Output
```



To change the bin width, first create a frequency distribution, then apply the histogram function:

```
      G←histogram 5 frequency #.SD.Height
      G.Output
```



1.3.4 Stem-and-Leaf Displays

A stem-and-leaf plot consists of two parts: the stem which is the leading digit(s) and the leaf which is the next digit to the right.

To create a stem-and-leaf display for the height data, simply enter:

```
      stemAndLeaf show #.SD.Height
5 | 5
6 | 1 1 2 2 4 5 5 6 6 7 7 7 8 9 9 9
7 | 0 0 0 1 1 1 1 1 1 1 2 2 2 2 2 3 3 4 4 5 5
```

Unfortunately, there are only 3 bins because they are 10 units wide. To create more bins, we divide them in half to create 5-unit bins. The optional left argument allows us to do this. First, we must determine the stem size which is equivalent to the bin width in a histogram. The stem size defaults 10. Thus, if we want to split the bins, we indicate this by a left argument of 5 units. Thus:

```
       5 stemAndLeafshow #.SD.Height
5 | 5
6 | 1 1 2 2 4
6 | 5 5 6 6 7 7 7 8 9 9 9
7 | 0 0 0 1 1 1 1 1 1 1 2 2 2 2 2 3 3 4 4
7 | 5 5
```

Now let’s try it for another variable: Weight.

If we choose a bin width of 100 pounds, we get too few bins:

```
      100 stemAndLeaf show #.SD.Weight
1 | 0 1 2 2 2 3 3 4 4 4 4 5 5 5 6 6 6 6 6 7 7 7 7 8 8 9 9 9
2 | 0 0 1 1 2 2 3 5 6 8
```

A bin width of 50 pounds is better, but we would like between 5 and 20 bins:

```
      50 stemAndLeaf show #.SD.Weight
1 | 0 1 2 2 2 3 3 4 4 4 4
1 | 5 5 5 6 6 6 6 6 7 7 7 7 8 8 9 9 9
2 | 0 0 1 1 2 2 3
2 | 5 6 8
```

Choosing a bin width of 10 lbs. gives us the following:

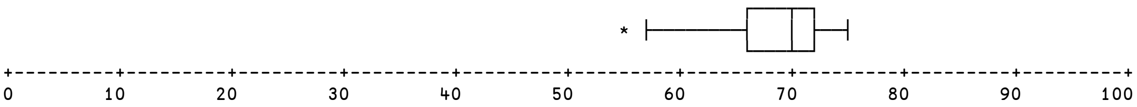
```
    10 stemAndLeaf show #.SD.Weight
10 | 0 5
11 | 5 5
12 | 0 5
13 | 0
14 | 0 0 0 2
15 | 0 0 0 5 5
16 | 0 0 0 5 5 5
17 | 0
18 | 0 0 5 5
19 | 0 5 5
20 | 5
21 | 0
22 | 0 0 5
23 |
24 | 5
25 |
26 | 0
27 |
28 | 0
```

Note that stems 23, 25 and 27 have no leaves, but the spacing is critical to show the shape of the data.

1.3.5 Box Plots

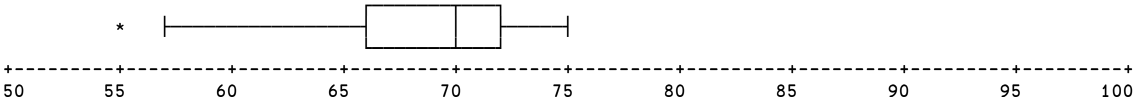
To create a box plot (also known as a box-and-whisker plot), we use the five number summary.

```
      boxPlot show #.SD.Height
```



There is a lot of white space to the left of the box plot. To create more detail, we can change the scale by starting at 50 inches:

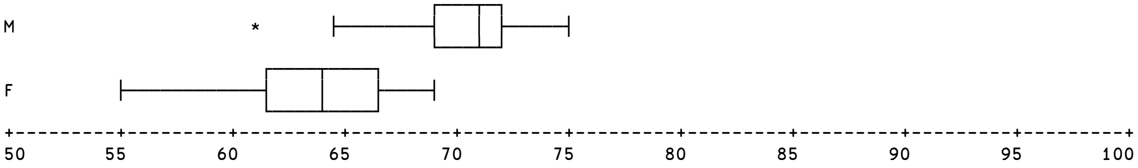
```
      50 boxPlot show  #.SD.Height
```



Notice the outlier at 55 is indicated by an asterisk (*).

Box plots are extremely useful when comparing groups. Suppose we would like to compare heights of male and female students.

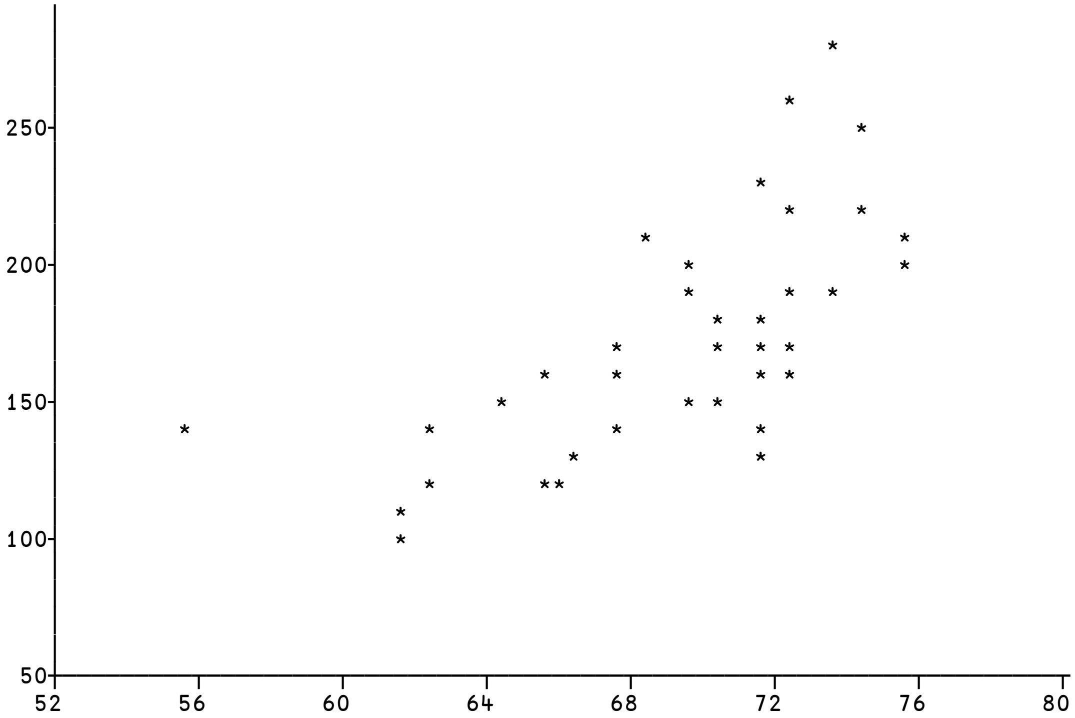
```
NS ← 50 boxPlot show #.SD.Height #.SD.Sex
NS.Output
```



1.3.7 Scatter Plots

To illustrate the relationship between two quantitative variables, we can create a scatter plot. The left argument contains the values for the y-axis, while the right argument supplies the values for the x-axis.

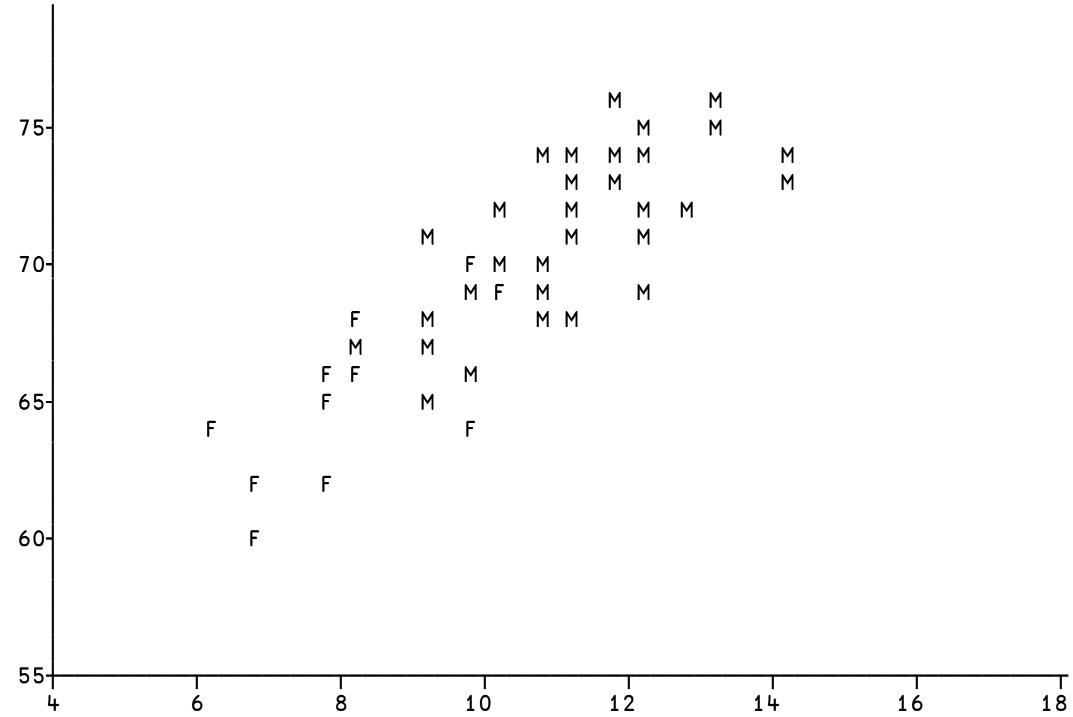
```
NS←scatterPlot #.SD.Weight #.SD.Height
NS.Output
```



The above plot shows that `Height` and `Weight` are moderately correlated, that the correlation is positive and since the points appear to cluster around a straight line, that the relationship is somewhat linear.

To differentiate between two or more groups, you can use a character field as the left argument:

```
NS←#.SD.Sex scatterPlot #.SD.(Height ShoeSize)
NS.Output
```



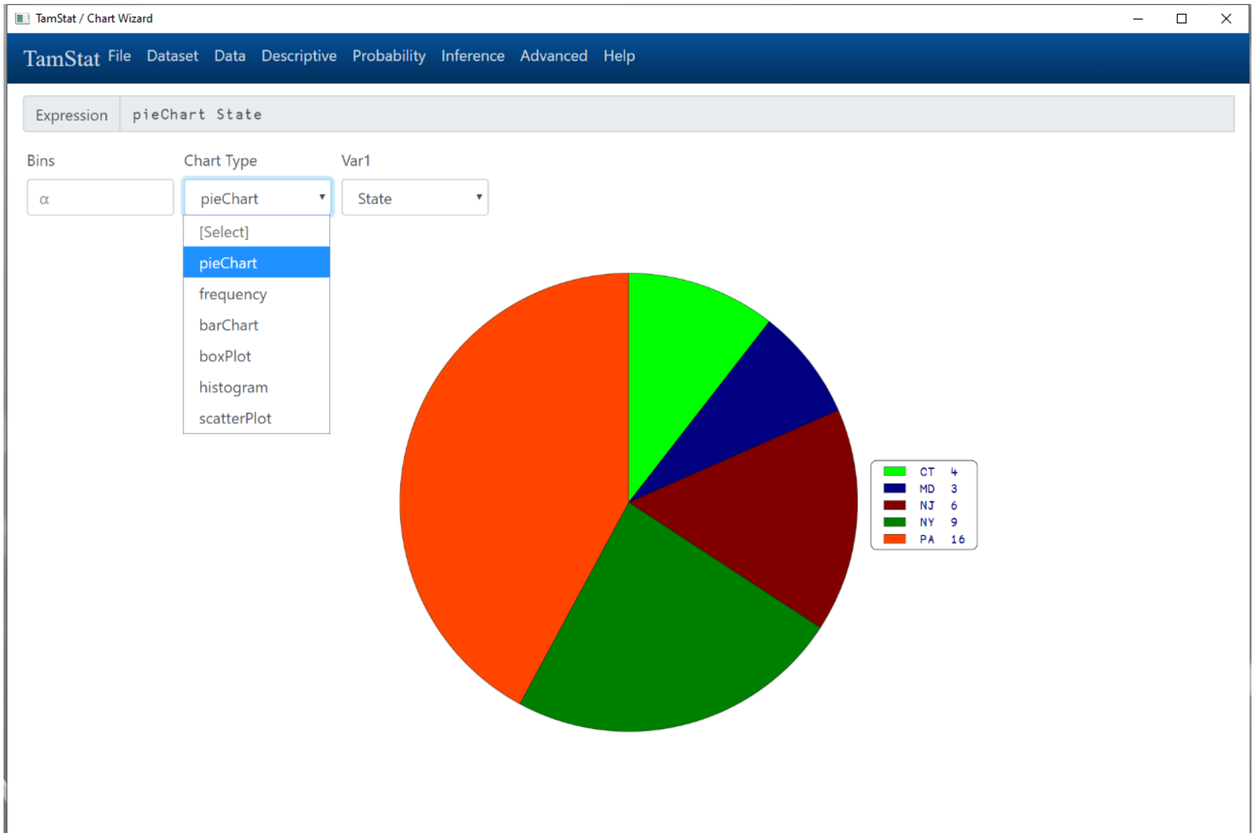
1.3.6 Dot Plots

To generate a dot plot, enter the following:

```
     dotPlot show #.SD.Height
                                                 *
                                                 *
                                                 *  *
                                                 *  *
                                     *     *  *  *  *
                   *  *        *  *  *     *  *  *  *  *  *  *
 *                 *  *     *  *  *  *  *  *  *  *  *  *  *  *
--------------------------------------------------------------
55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
```

1.3.7 Graphics Wizard

Although graphics can be created using the session, users can also create graphics using the Graphics Wizard. In the desktop version, select Graphics from the main menu; in the HTML version, select Descriptive, then Graphics. The user then has the option to select the chart type, one or two variable names and an optional grouping:

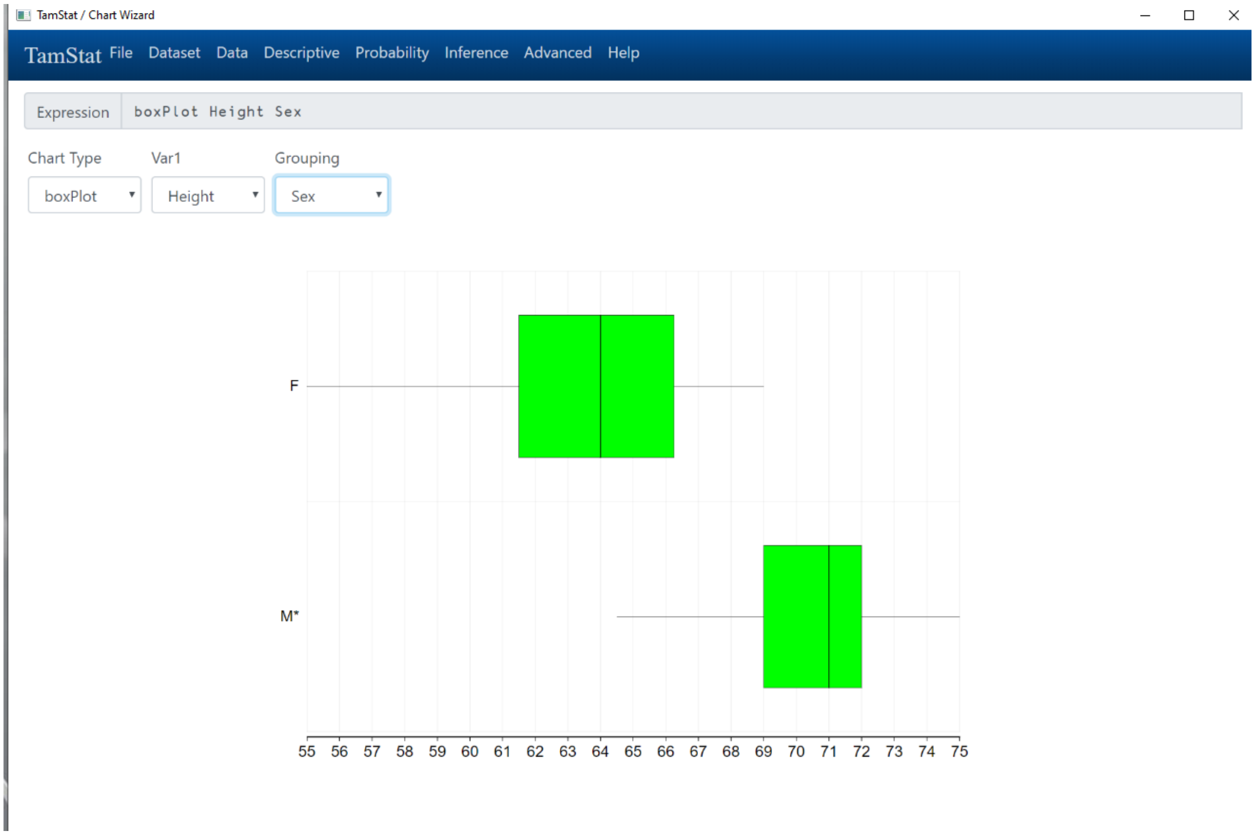


Expression `pieChart State`

Bins: α

Chart Type: pieChart

Var1: State



Expression `boxPlot Height Sex`

Chart Type: boxPlot

Var1: Height

Grouping: Sex

1.4 Obtaining Summary Statistics from tabular data

While it is useful to obtain statistics from raw data, it may also be useful to obtain statistics from tabular data. Let's obtain the relative frequency of family size from the student data. When the second column of a frequency table contains positive integers, the frequency distribution represents a sample because the second column contains count data:

```
      SAMPLE_FREQ←frequency #.SD.Family
      SAMPLE_FREQ
0  2
1 17
2 11
3  7
4  1
      mean SAMPLE_FREQ
1.6842
      var SAMPLE_FREQ
0.87055
```

A frequency table represents a population when its second column sums to 1 and contains values strictly between zero and 1.

```
     FR←relative frequency #.SD.Family
0 0.052632
1 0.44737
2 0.28947
3 0.18421
4 0.026316
     mean FR
1.6842
     var FR
0.84765
```

Note that while the sample and population means are the same, the variances are slightly different. This is because the sample variance uses the denominator *n*-1.

When we calculate frequencies using ranges, the results may be slightly different. For example the frequency function produces a table. The table does not indicate that there are 3 people whose height is 63. It indicates that there are 3 people whose height is between 62 and 64 inches. The indicator is the midpoint of the range (not the lower bound), to eliminate bias in calculations.

```
     frequency #.SD.Height
54  1
57  0
60  2
63  3
66  7
69  7
72 14
75  4
```

Note that the mean calculation for the raw data is slightly different from that of the grouped data. This is because there is some loss of information in the latter case:

```
      mean #.SD.Height
68.77631579
      mean frequency #.SD.Height
68.84210526 1.5 Data Selection
```

Sometimes you need to exclude certain values from a sample. This may include outliers are data that are unavailable. Here we have an example of ACT Test scores:

```
      ACT
22 20 31 28 27 25 20 27 24 21 19 24 24 34 0 19 22 0 24 30
      mean ACT
22.05
```

Suppose we wanted to exclude ACT scores over 30. We could do the following:

```
      ACT excluding > 30
22 20 28 27 25 20 27 24 21 19 24 24 0 19 22 0 24 30
      mean ACT excluding > 30
20.88888889
```

We may want to find any outliers in the data set. It turns out there are two:

```
      outliers ACT
0 0
```

The zeroes may represent missing data and may distort the mean. To exclude them, we can do the following:

```
      mean ACT excluding outliers ACT
24.5
```

Note that the mean is about 2.5 points higher without the outliers.

To exclude specific values:

```
    ACT excluding = 20
22 31 28 27 25 27 24 21 19 24 24 34 0 19 22 0 24 30
    ACT excluding 0,20
22 31 28 27 25 27 24 21 19 24 24 34 19 22 24 30
    ACT excluding in 0 20
22 31 28 27 25 27 24 21 19 24 24 34 19 22 24 30
```

To split a data set into groups, use the `splitBy` function. The first group is the selected group; the second group is everything else:

```
      MALE FEMALE ← #.SD.Height splitBy #.SD.Sex eq 'M'
      MALE
72 69 69 70 66 67 72 72 71 71 73 65 73 67 70 68 74 71 72 75 70 71 71 71 74 61 75 72 71
      FEMALE
62 61 69 65 55 67 64 62 65.5
```

To perform queries on data, we can use “`selectFrom` with “`where`”

Select the heights from students in Pennsylvania:

```
       selectFrom #.SD.Height where #.SD.State in 'PA'
72 62 69 72 73 61 65 73 70 71 71 71 61 67 64 71
```

Select the heights of female students from a database:

```
      'Height' selectFrom D where #.SD.Sex eq 'F'
62 61 69 65 55 67 64 62 65
```

Create a new database of students from Pennsylvania:

```
     DB←selectFrom D where #.SD.State in 'PA'
     DB.Family
3 4 1 2 3 3 3 1 1 2 1 2 2 1 2 1
     DB.Height
72 62 69 72 73 61 65 73 70 71 71 71 61 67 64 71
```

Create a new database consisting of heights and shoe sizes of female students:

```
      DF← 'Height,ShoeSize' selectFrom D where #.SD.Sex eq 'F'
      DF.Height
62 61 69 65 55 67 64 62 65.5
      DF.ShoeSize
9 6.5 9.5 7.5 8.5 9.5 8 6 7
```

1.6 Exercises

1. Using the STATE variable from the class data set:
   a. Create a frequency distribution
   b. Create a pie chart and a bar chart
   c. Most students are from what state?
   d. What percentage of students are from that state?

2. Create a contingency table of STATE versus SEX
   a. How many male students are from Pennsylvania?
   b. What proportion of students from Pennsylvania are male?
   c. What proportion of female students are from New Jersey?

3. Using the class data set :

   a. Create a histogram and a stem-and-leaf plot of student weights. Describe the shape of the distribution.
   b. Find the mean, median and standard deviation. Is the mean greater than, less than or approximately equal to the median?
   c. Break down the above data by sex. Compare the means and standard deviations
   d. Create a box plot of male weight vs. female weight. Does the graphic agree with your conclusions in part c.?

4. Create a frequency distribution for number of siblings (Use `#.SD.Family`).
   a. How many students have no brothers and sisters?
   b. How many students have no more than one sibling?
   c. What percentage of students have no more than one sibling?
   d. Create a bar chart (or histogram) and describe the shape of the data.
   e. Find the skewness. Is it positive, negative or approximately zero? Does this agree with you conclusion in part d?

Chapter 2 – Probability

There are three types of probability: Theoretical, empirical and subjective. In this section we will look at empirical probability by examining the rules of probability using a data set and at theoretical probability by looking at various probability distributions.

Probabilities range from 0 (impossible) to 1 (certain). Probabilities greater than ½ are considered likely or probable, whereas probabilities less than ½ are considered possible. Probabilities less than 0.05 are considered unlikely.

A **trial** is a simple experiment such as tossing a coin or rolling a die and observing the result.

An **outcome** is the result of a trial. The outcome of tossing a coin is either a head (H) or a tail (T). The outcome of rolling a die is one of the numbers between one and six.

An **event** is a set of outcomes. Rolling greater than 4 on a die is an event consisting of the two outcomes 5 and 6. A simple event consists of a single outcome; thus tossing a coin and getting a head is a simple event.

The set of all events is known as the **sample space**. The sample space for tossing a coin is {H,T} whereas the sample space for rolling a die is {1,2,3,4,5,6}.

The occurrence of an event is a **success**; the non-occurrence of an event is a **failure**. We represent a success by the number 1 and a failure by 0.

2.1 Theoretical Probability

We define theoretical **probability** as the number of successes divided by the total number of outcomes. The assumption is that each outcome is equally likely.

For a coin toss, there are two possible outcomes: heads and tails. Only one outcome, Heads, is a success. Therefore the probability of heads is one success divided by two outcomes or 1/2:

```
    X←'H,T'
    proportion X eq 'H'
0.5
```

The probability that X>4 on the roll of a die can be calculated by dividing the number of successful outcomes (2) by the number of outcomes in our sample space (6). Thus the probability is 1/3.

```
    X←1 2 3 4 5 6
    X>4
0 0 0 0 1 1
    proportion X>4
0.33333
```

In a deck of cards there are four suits: Spades, Hearts, Diamonds and Clubs and 13 ranks: Ace,King,Queen,Jack,10,9,8,7,6,5,4,3 and 2.

Let us create two variables: Suit and Rank:

```
     RANK←Y,Y,Y,Y←toNestedVector 'A,K,Q,J,10,9,8,7,6,5,4,3,2'
     SUIT←13 rep toNestedVector 'Spade,Heart,Diamond,Club'
```

In blackjack or 21, tens and face cards are worth 10 points each. The probability that a card is worth 10 points can be calculated by selecting the number of tens and face cards. There are four suits and three face cards in each suit

(King, Queen and Jack). So, there are 12 face cards and four tens for a total of 16 successes. There are 52 cards in the deck, so the probability is:

```
    16 div 52
0.30769

    proportion RANK in 'K,Q,J,10'
0.30769
```

2.2 Empirical Probability

Empirical probability is based on observed data. The probability of an event is calculated by the number of successes in the same (x) divided by the sample size (n). From the student data we obtain the following frequency distribution:

```
    frequency #.SD.Sex
 F   9
 M  29
```

The probability that a student is female is calculated by taking the number of female students (9) divided by the total number of students in the sample (9+29=38). Thus

```
    proportion #.SD.Sex eq 'F'
0.23684
```

2.3 Rules of Probability

There are four rules of probability: the complement rule, the addition rule, the multiplication rule and the conditional rule. They roughly correspond to the four logical functions: not, and, or and if as well as the four arithmetic operations: -,×,+,÷

Rules of Probability (Summary)					
Term	Symbol	Condition	Special Formula	General Formula	Primary Operation
Complement (not A)	A'	None	$P(A') = 1 - P(A)$		−
Union (A or B)	A∪B	Mutually Exclusive	P(A∪B) = P(A) + P(B)	P(A∪B) = P(A) + P(B) - P(A∩B)	+

Intersection (A and B)	A∩B	Independent	P(A∩B) = P(A)P(B)	P(A∩B) = P(A)P(B|A)
				P(A∩B) = P(A) + P(B) - P(A∪B)
					×

Conditional (A if B)	A|B	Independent	P(A|B) = P(A)	P(A|B) = P(A∩B)/P(B)	÷
		Mutually Exclusive	P(A|B) = 0		

2.3.1 The Complement Rule

The complement of an event is the non-occurrence of an event. Since an event must either occur or not occur, the probability of non-occurrence is one minus the probability of the event.

From the student data we obtain the following frequency distribution showing that there are 11 democrats, 11 independents and 16 republicans out of the 38 students:

```
    frequency #.SD.Party
 D  11
 I  11
 R  16
```

The probability that a randomly selected student is a Republican is:

```
   #.SD.Party prob eq 'R'
0.42105
```

The complement is the event that a student is not a Republican.   The probability is 1-0.4205 or:

```
   #.SD.Party prob ne 'R'    ⍝ Probability random student is not Republican
0.5789473684
```

The probability that a card drawn at random is not a spade is

```
    proportion SUIT ne 'Spade'
0.75
```

2.3.2 The Multiplication Rule

Suppose we want to find the probability of tossing two coins and finding them both heads.   The probability of heads on the first coin is ½, and the probability of heads on the second coin is ½. The sample space consists of four events {HH,HT,TH,TT}. Only one of those contains both heads. So the probability is ¼.

```
    proportion 'HH,HT,TH,TT' eq 'HH'
0.25
```

We can also use the multiplication rule to find this.   The probability of heads on both coins is equal to the probability of heads on the first coin times the probability of heads on the second coin.

```
    0.5(and prob independent).5
0.25
```

What is the probability that a card drawn from a deck of 52 is a red ace?

```
    proportion (RANK eq 'A') and(SUIT in 'Heart,Diamond')
0.038462
    ACE ← 1÷13     ⍝ Probability of an Ace
    RED ← 26÷52    ⍝ Probability of a red card = 0.5
    ACE (and prob independent) RED
0.038462
```

Again, we can show using the multiplication rule that:

```
    (proportion RANK eq 'A') times (proportion SUIT in 'Heart,Diamond')
0.038462
```

From the student data, let us create a contingency table using the variables Sex and Party.   The Variable #.SD.Sex consists of the values M=Male and F=Female. The variable #.SD.Party consists of the values D=Democrat, I=Independent and R=Republican.

```
    Table←frequency #.SD.Sex #.SD.Party
    D  I   R
 F  3  2   4
 M  8  9  12
```

What is the probability that a student is both a republican and male? From the data we can see that there are 8 students out of 38 who are both republican and male. Thus, the probability is 12/38 or:

```
    proportion (#.SD.Party eq 'R') and (#.SD.Sex eq 'M')
0.31579
     'R' (and prob Table)'M'
0.3157894737
```

Now let's apply the multiplication rule:

```
      #.SD.Party prob eq 'R'
0.42105
      #.SD.Sex prob eq 'M'
0.76316
      .42105 times .76316
0.32133
```

Hey, what gives? The multiplication rule does not agree with the counting rule. What is different from the first example where this worked? The multiplication rule only works when the two events are **independent** of each other. In the first example, the first coin toss has no effect on the second. In the second example, the variables `Sex` and `Party` are not independent of each other. That is, the probability that a person is male or female is affected by his or her political affiliation.

2.3.3 The Addition Rule

Now let us find the probability that a student is a Democrat or a Republican. We can simply add the probabilities:

```
     proportion #.SD.Party eq 'D'
0.28947
     proportion #.SD.Party eq 'R'
0.42105
     proportion (#.SD.Party eq 'D') or (#.SD.Party eq 'R')
0.71053
     'D' (or prob Table)'R'
0.7105263158
```

What is the probability that a student is either Republican or male? Let's try the addition rule:

```
     R←proportion #.SD.Party eq 'R'
0.42105
     M←proportion #.SD.Sex eq 'M'
0.76316
     R+M
1.1842
```

How can we have a probability that is greater than one? We actually counted the students that were both male and republican twice. Let's look at the contingency table again:

```
      Table
       D    I   R  Total
 F      3    2   4     9
 M      8    9  12    29
Total  11   11  16    38
```

There are a total of 16 Republicans and 29 male students. The students who are Republican or male are:

```
    8+9+12+4
33
```

So, the probability that a student is either Republican or male is: 33/38.

```
    proportion (#.SD.Party eq 'R')or(#.SD.Sex eq 'M')
0.86842
    'R' (or prob Table)'M'
0.86842
```

To compensate for the double counting, we must subtract the probability that both events occur:

```
    RANDM←proportion (#.SD.Party eq 'R') and (#.SD.Sex eq 'M')
0.31579
    RANDM<-'R' (and prob Table)'M'
0.31579
    R+M-RANDM
0.86842
```

So why did we not have to compensate for double counting in the first example? That is because the two events D and R were mutually exclusive; they could not both occur at the same time. No one can be both a Democrat and a Republican. Notice that the probability of both events occurring is 0:

```
    proportion (#.SD.Party eq 'D') and (#.SD.Party eq 'R')
0
    'D' (and prob Table)'R'
0
```

We could still apply the special rule, but we don’t have to worry about double counting a group that has no members.

In another example, what is the probability that a card is an ace or a spade? First, we find the marginal probabilities:

```
     A←proportion RANK eq 'A'
0.076923
     S←proportion SUIT eq 'Spade'
0.25
```

Then we find the joint probability:

```
     AS←proportion (RANK eq 'A')and (SUIT eq 'Spade')
0.019231
     ACE←1÷13
     SPADE←1÷4
     AS←ACE(and prob independent)SPADE
0.01923076923
```

Finally, we add the marginal probabilities and subtract the joint probability:

```
     ACE+SPADE-AS
0.30769
```

This gives us the correct result for the union of two groups, eliminating the double counting:

```
     proportion (RANK eq 'A') or (SUIT eq 'Spade')
0.30769
     ACE(or prob independent)SPADE
0.3076923077
```

2.3.3 The Conditional Rule

Earlier we mentioned that the variables Sex and Party are not independent of each other, because knowing something about one variable gives us information about the other. Not having any other information, we can find the probability that a student is a Republican:

```
      proportion #.SD.Party eq 'R'
0.42105
```

Suppose we randomly selected a male student. What is the probability he is republican? We can simply take the number of students who are both male and Republican, and divide by the number of male students. So the probability is now 12 out of 29 or:

```
      proportion (#.SD.Party eq 'R') given (#.SD.Sex eq 'M')
0.41379
      'R' (| prob Table)'M'
0.4137931034
```

Suppose we randomly selected a republican student. What is the probability that the student is male? Surprisingly this probability is not the same. Again we take the 12 students that are both male and republican and divide by the total number of Republicans; the probability is 12 out of 16:

```
      proportion (#.SD.Sex eq 'M') given (#.SD.Party eq 'R')
0.75
      'M' (| prob Table)'R'
0.75
      'M' (| prob #.SD.Sex #.SD.Party)'R'
0.75
```

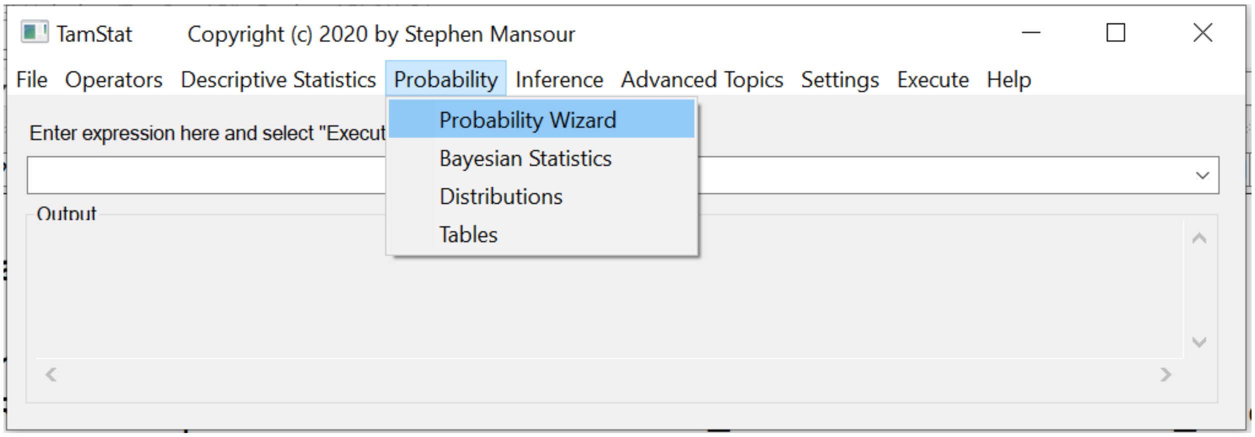
Since the conditional probability is different than the original probability, we can conclude that Sex and Party are not independent.

In the deck of cards, the rank is independent of the suit because the conditional probability is the same as the marginal probability:

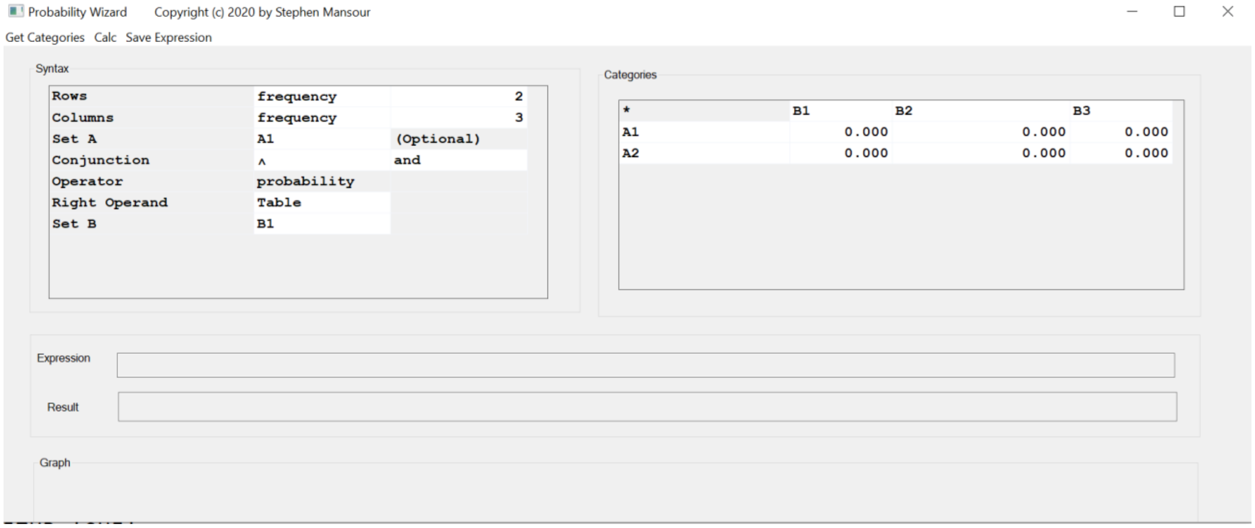
```
      proportion (RANK eq 'A') given (SUIT eq 'Spade')
0.076923
      proportion RANK eq 'A'
0.076923
       (1÷13)(|prob independent)0.25
0.07692307692
       (1÷13)(¬prob independent)0.25
0.07692307692
```

2.3.4 Probability Wizard

To use the Probability Wizard, select **Probability**, then **Probability Wizard**:



The following screen will appear:



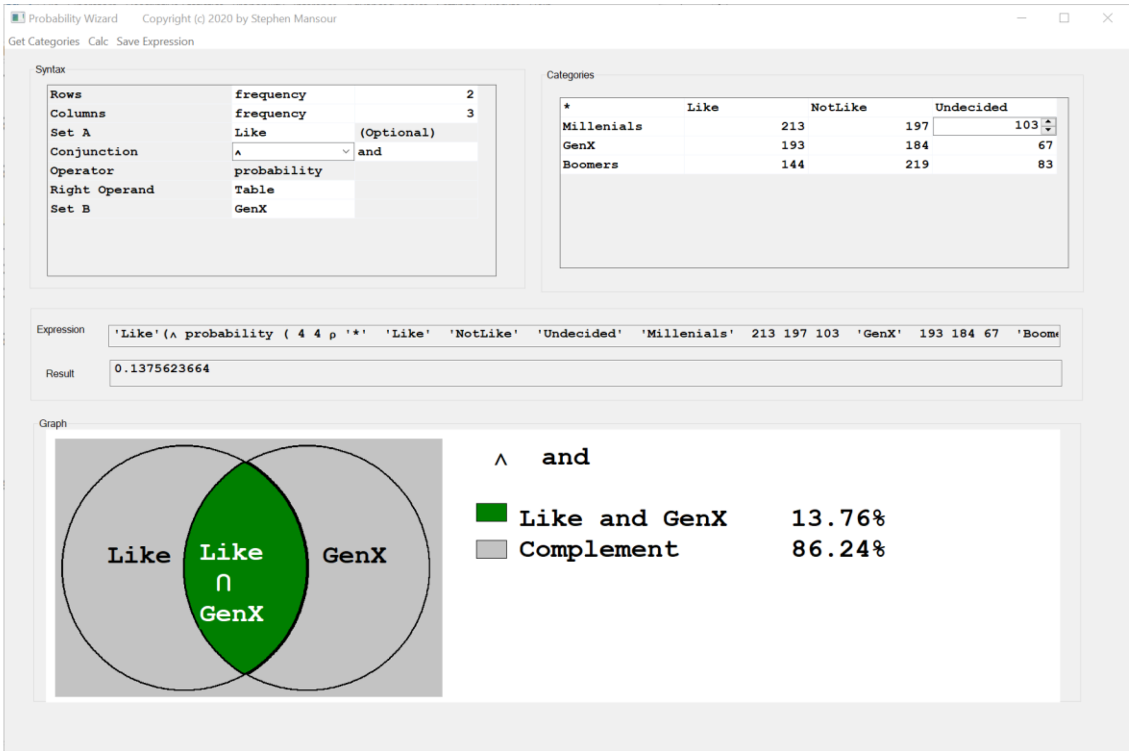
Rows	frequency	2
Columns	frequency	3
Set A	A1	(Optional)
Conjunction	∧	and
Operator	probability	
Right Operand	Table	
Set B	B1	

*	B1	B2	B3
A1	0.000	0.000	0.000
A2	0.000	0.000	0.000

To calculate probabilities using summary data, enter the number of rows and columns in the upper right corner of the syntax group. Then enter the summary data in the categories group. First enter the row and column titles, then enter the data in the body of the table. The data may be counts, probabilities or percents. Then select two subsets (corresponding to the row and/or column names in the categories groups and enter them in SetA and SetB. Finally select a conjunction by name or by symbol. Examples are listed on the next page.
We are interested in GenXers who like a particular product being marketed.

$$P(Like|GenX) = \frac{P(Like \cap GenX)}{P(GenX)}$$

To find the probability of selecting a GenXer who likes the product:



Probability Wizard   Copyright (c) 2020 by Stephen Mansour

Get Categories Calc Save Expression

Syntax

Rows	frequency		2
Columns	frequency		3
Set A	Like	(Optional)	
Conjunction	∧	and	
Operator	probability		
Right Operand	Table		
Set B	GenX		

Categories

*	Like	NotLike	Undecided
Millenials	213	197	103
GenX	193	184	67
Boomers	144	219	83

Expression `'Like'(∧ probability ( 4 4 ρ '*'  'Like'  'NotLike'  'Undecided'  'Millenials'  213 197 103  'GenX'  193 184 67  'Boome`

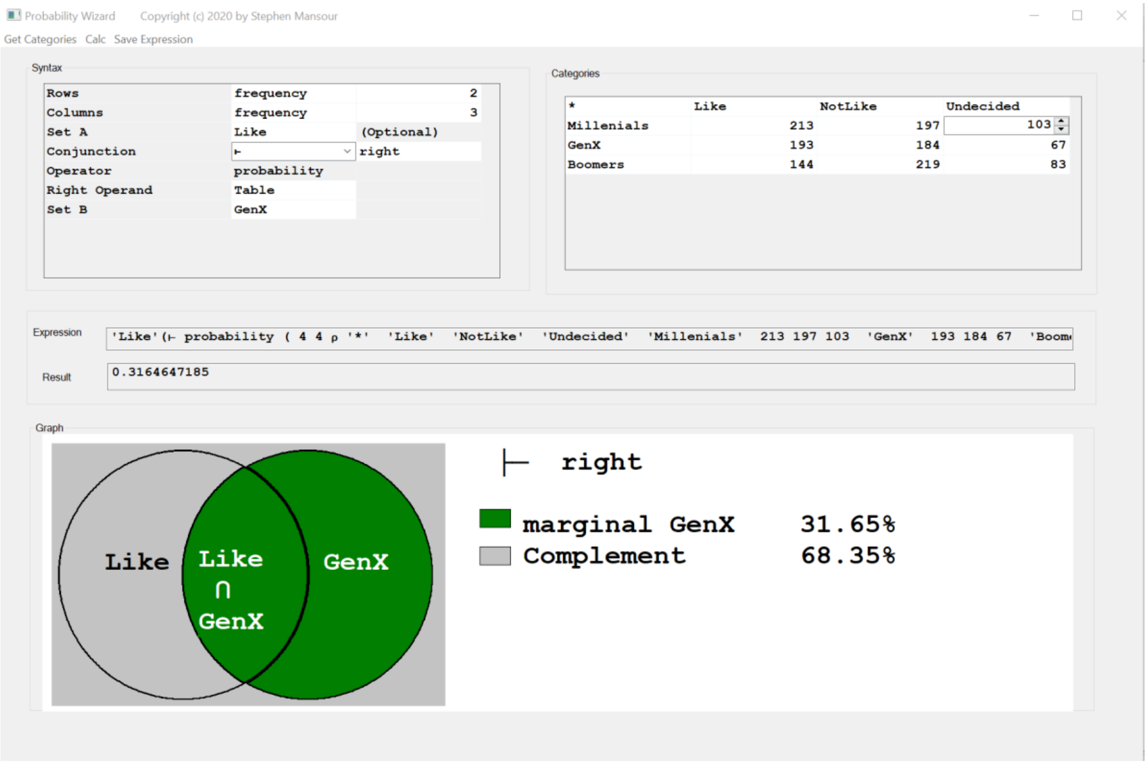
Result `0.1375623664`

Graph

∧ and

🟩	Like and GenX	13.76%
▨	Complement	86.24%

The marginal probability of selecting GenX:



Probability Wizard   Copyright (c) 2020 by Stephen Mansour

Get Categories Calc Save Expression

Syntax

Rows	frequency		2
Columns	frequency		3
Set A	Like	(Optional)	
Conjunction	⊢	right	
Operator	probability		
Right Operand	Table		
Set B	GenX		

Categories

*	Like	NotLike	Undecided
Millenials	213	197	103
GenX	193	184	67
Boomers	144	219	83

Expression `'Like'(⊢ probability ( 4 4 ρ '*'  'Like'  'NotLike'  'Undecided'  'Millenials'  213 197 103  'GenX'  193 184 67  'Boom`

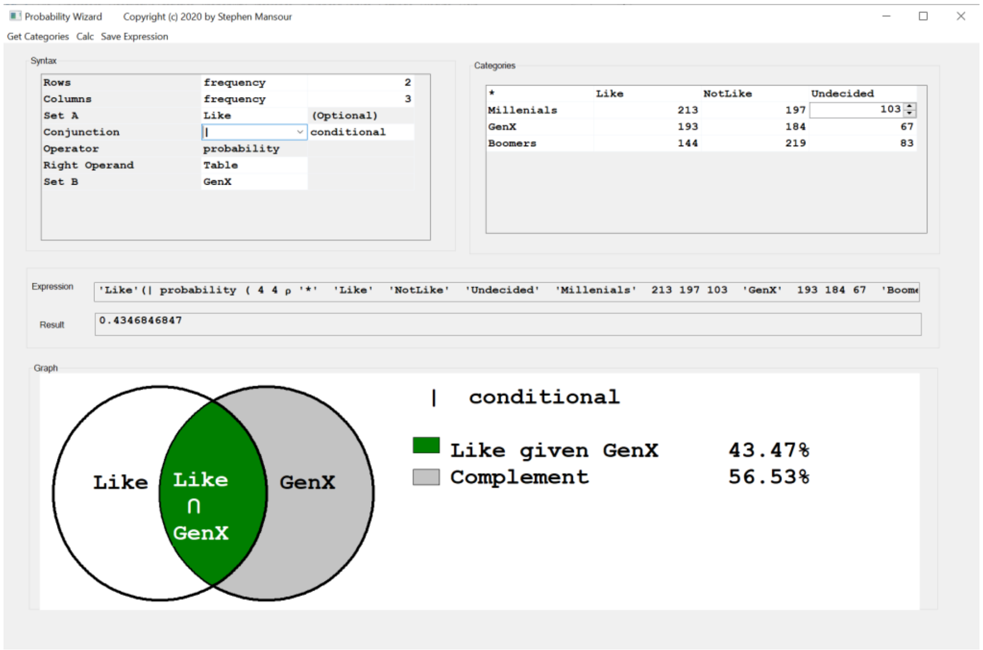
Result `0.3164647185`

Graph

⊢ right

🟩	marginal GenX	31.65%
▨	Complement	68.35%

Conditional probability that a GenXer will like the product:



Rows	frequency	2
Columns	frequency	3
Set A	Like	(Optional)
Conjunction	|	conditional
Operator	probability	
Right Operand	Table	
Set B	GenX	

*	Like	NotLike	Undecided
Millenials	213	197	103
GenX	193	184	67
Boomers	144	219	83

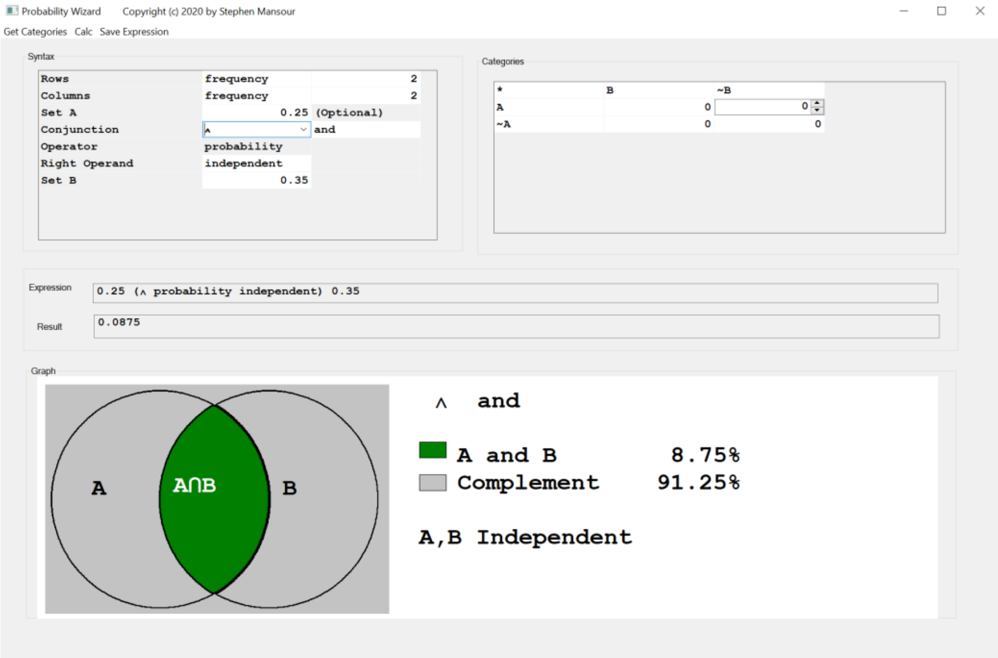
Expression `'Like'(| probability ( 4 4 ρ '*'  'Like'  'NotLike'  'Undecided'  'Millenials'  213 197 103  'GenX'  193 184 67  'Boome`

Result `0.4346846847`

| conditional

🟩 Like given GenX 43.47%
▨ Complement 56.53%

To calculate independent probabilities, select ‘independent’ as the right operand and enter the probabilities for SetA and SetB. If the conjunction “and” is selected, the joint probability is simply the product of the two individual probabilities. Probabilities involving different conjunctions can be calculated from the joint probability and the individual probabilities.



Rows	frequency	2
Columns	frequency	2
Set A	0.25	(Optional)
Conjunction	∧	and
Operator	probability	
Right Operand	independent	
Set B	0.35	

*	B	~B
A	0	0
~A	0	0

Expression `0.25 (∧ probability independent) 0.35`

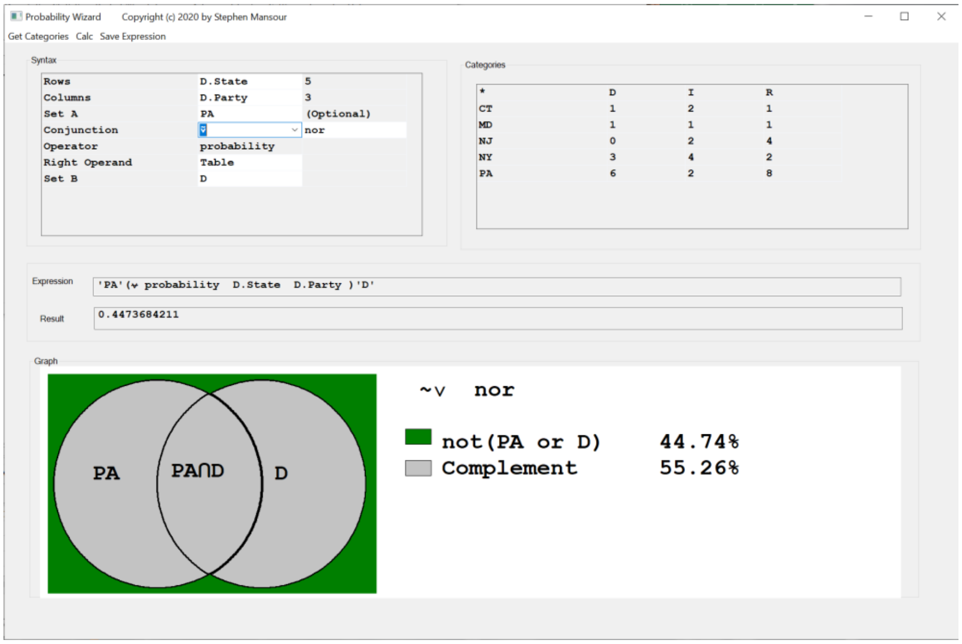
Result `0.0875`

∧ and

🟩 A and B 8.75%
▨ Complement 91.25%

A,B Independent

Users can also obtain probabilities from a database. To determine if a randomly selected students is neither from PA nor a Democrat, select the fields “State” and “Party” from the database for Rows and Columns using the dropdowns. Select “nor” for the conjunction, then select “PA” and “D” for SetA and SetB respectively. When users select “Calc”, a read-only contingency table will display in the “Categories” Group, in addition to the the expression and result. The screen is displayed below:


Probability Wizard   Copyright (c) 2020 by Stephen Mansour
Get Categories Calc Save Expression
Syntax
Categories
Expression
Result
Graph
PA
PA∩D
D

Rows	D.State	5
Columns	D.Party	3
Set A	PA	(Optional)
Conjunction	⍱	nor
Operator	probability	
Right Operand	Table	
Set B	D	

*	D	I	R
CT	1	2	1
MD	1	1	1
NJ	0	2	4
NY	3	4	2
PA	6	2	8

```
'PA'(⍱ probability  D.State  D.Party )'D'
```

```
0.4473684211
```

~⍱ nor

🟩 not(PA or D)   44.74%
▢ Complement   55.26%

2.4 Bayesian Statistics

The conditional probability rule can be used in different ways. Suppose the probability that a person has cancer is 3%. A certain test will be positive 90% of the time when a person has cancer. But there is a 2% chance of a false positive. What is the probability that a person actually has the disease if the result is positive? We know P(Test|Disease) = 0.9. What we are trying to find is P(Disease|Test).

We use the conditional rule: $\mathrm{P}(\mathrm{A}|\mathrm{B}) = \mathrm{P}(\mathrm{A} \cap \mathrm{B})/\mathrm{P}(\mathrm{B})$ and P(B|A) = P(A∩B)/P(A).

From this we can show:

$$\mathrm{P}(\mathrm{A}|\mathrm{B})\mathrm{P}(\mathrm{B}) = \mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \mathrm{P}(\mathrm{B}|\mathrm{A})\mathrm{P}(\mathrm{A})$$

We can then show that P(B|A) = P(A|B)P(B)/P(A).

The marginal probability $P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)\, P(B_i)$

So we can show:

$$P(B_i|A) = \frac{\mathrm{P}(A|B_i)\mathrm{P}(B_i)}{\sum_i P(A|B_i)\, P(B_i)}$$

Let us first set the prior probabilities: P(Cancer), P(No Cancer)

```
      PRIOR←0.03 0.97
```

Now let us set the conditional probabilities: P(Positive|Cancer) , P(Postive|No Cancer)

```
      COND←0.9 0.02
```

Now let find the Bayesian probabilities: P(Cancer|Positive),P(No Cancer|Positive)

```
      bayes PRIOR COND
0.5819 0.4181
```

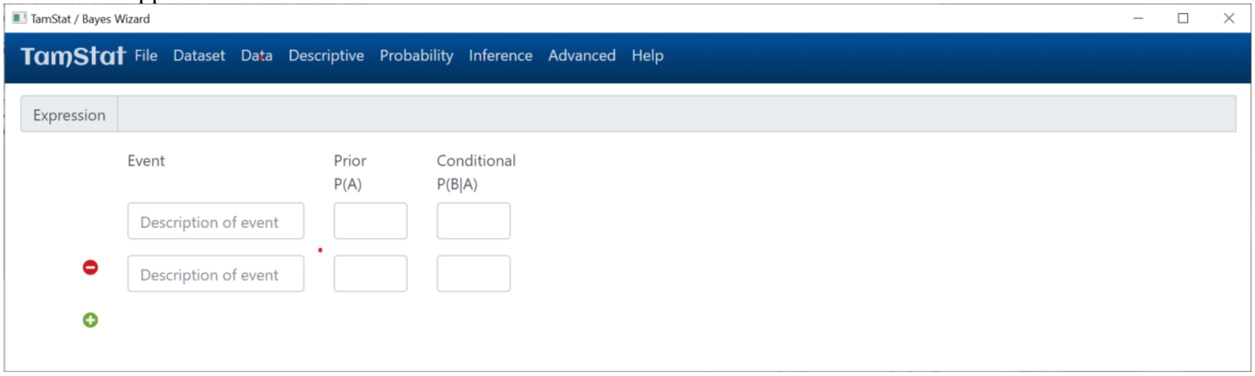
To obtain the details, use the show operator:

```
       'Cancer,No Cancer' bayes show PRIOR COND
Event          Prior     Cond   Joint     Posterior
Cancer         0.03       0.9   0.027     0.5819
No Cancer      0.97       0.02  0.0194    0.4181
Total          1      Marginal  0.0464    1
```

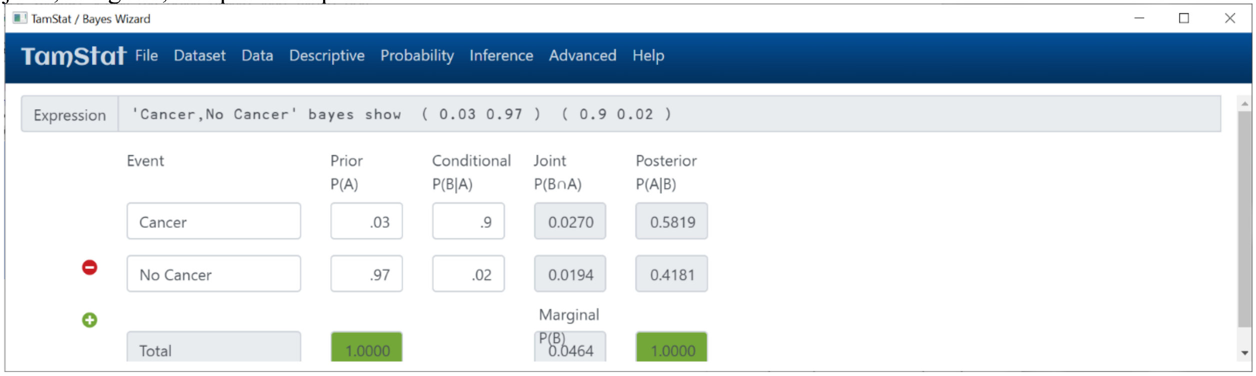
Thus, in conclusion we find that if the test is positive, the probability of cancer is 58.19%.

2.4.1 Bayesian Wizard

From the main menu, select "`Advanced Topics`", then select "`Bayesian Statistics`". The following screen will appear:



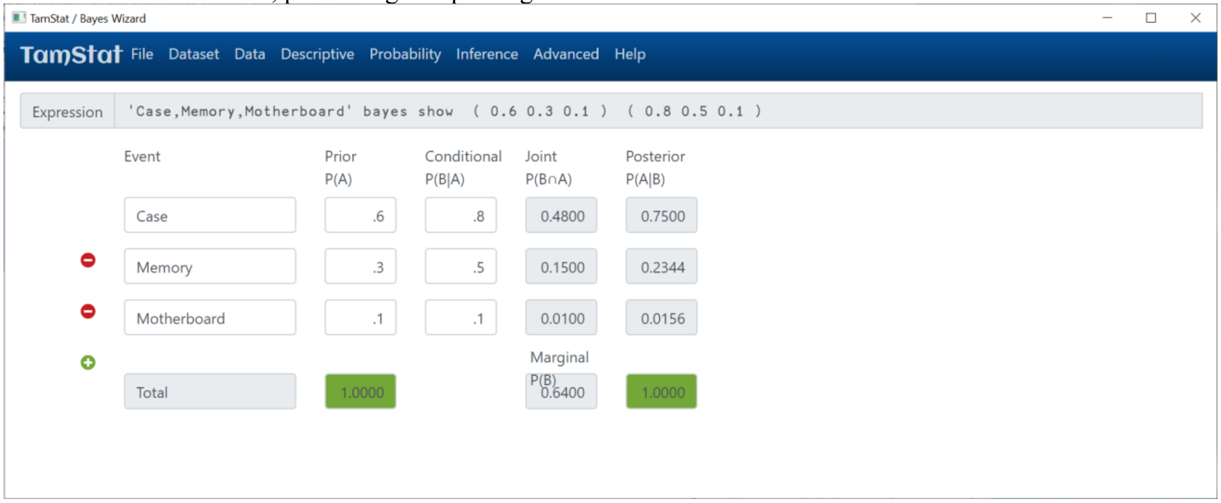
Enter data into the first three columns: Event name, prior and conditional probabilities. TamStat will calculate the joint, marginal, and posterior probabilities.


TamStat / Bayes Wizard
TamStat File Dataset Data Descriptive Probability Inference Advanced Help

Expression `'Cancer,No Cancer' bayes show ( 0.03 0.97 ) ( 0.9 0.02 )`

Event	Prior P(A)	Conditional P(B|A)	Joint P(B∩A)	Posterior P(A|B)
Cancer	.03	.9	0.0270	0.5819
No Cancer	.97	.02	0.0194	0.4181
			Marginal P(B)	
Total	1.0000		0.0464	1.0000

To insert additional rows, press the green plus sign.


TamStat / Bayes Wizard
TamStat File Dataset Data Descriptive Probability Inference Advanced Help

Expression `'Case,Memory,Motherboard' bayes show ( 0.6 0.3 0.1 ) ( 0.8 0.5 0.1 )`

Event	Prior P(A)	Conditional P(B|A)	Joint P(B∩A)	Posterior P(A|B)
Case	.6	.8	0.4800	0.7500
Memory	.3	.5	0.1500	0.2344
Motherboard	.1	.1	0.0100	0.0156
			Marginal P(B)	
Total	1.0000		0.6400	1.0000

2.5 Counting Rules

In theoretical probability, the probability of an event is the number of successful outcomes divided by the total number of outcomes. Thus, in order to compute theoretical probabilities, one must learn how to count.

2.5.1 Multi-Step Experiment

Identical Trials. If we perform the same experiment more than once, how many possible outcomes are there? The formula for performing k trials, each having *n* outcomes is:

$$k^n$$

Flipping a coin 7 times. There are 7 trials and 2 outcomes (heads and tails) for each trial.

```
      2*7     ⍝ In TamStat we use the power function:    K*N
128
```

Throwing 5 dice (as in Yahtzee). There are 5 dice and 6 outcomes for each die:

```
      6*5
7776
```

Non-Identical Trials. If the probabilities differ for each experiment, the general formula for k experiments is : $n_1 \times n_2 \times \cdots \times n_k$ where n_i is the number of outcomes for experiment *i*.

Tossing a coin, then throwing a die. There are 2 outcomes for the first trial and 6 for the second; thus

```
      2 × 6
12
```

In Pennsylvania, standard license plates consist of 3 letters followed for 4 digits, e.g. ABC1234. How many possible license plates are there? . There are 26 outcomes for each of the first three symbols, plus 10 outcomes for each of the last four. The result is: $26 \times 26 \times 26 \times 10 \times 10 \times 10 \times 10 = 26^3 10^4$

```
      (26*3) × (10*4)
175760000
```

The last PA census was only 12,281,284 so there are plenty of remaining license plate combinations.

2.5.2 Factorials, Permutations and Combinations

Factorials. How many ways can one arrange three items without repeating A,B,C? There are 3 possibilities for the first item, two for the second and one for the third $3 \times 2 \times 1 = 6$. The number of ways to arrange n objects is $n!$

$$n! = n \times (n-1) \times \cdots \times 2 \times 1$$

There are six ways to arrange 3 objects: {ABC,ACB,BAC,BCA,CAB,CBA}

```
      !3
6
```

How many ways can one arrange 7 objects? 10 objects?

```
   !7 10
5040 3628800
```

One can see that factorials get surprisingly large very quickly. That is why we use the exclamation point as the symbol for this function.

Permutations. Suppose there is an Olympic event in which there are 10 contestants. How many possible results are there if we are only interested in the three medalists? Using the multi-step experiment approach, there are 10 possibilities for the gold medal, nine remaining for the silver, and 8 for the bronze. Thus there are 10x9x8 = 720 possibilities. Writing this in terms of factorials :

$${}_{10}P_3 = \frac{10!}{(10-3)!} = \frac{10!}{7!} = 10 \times 9 \times 8$$

```
      10×9×8
720
      (!10)÷(!7)
720
```

Or more succinctly:

```
      ÷/!10 7
720
```

Combinations. Let's look at a similar problem. Suppose we need to hire 3 bank tellers and we get 10 job applications. The only thing that matters is who gets hired. Order does not matter. From the previous problem we know that there are 720 ways to accomplish this when order is important. Since there are 3!=6 ways to arrange 3 objects, we simply divide 720 by 6 = 120.

$${}_{10}C_3 = \binom{10}{3} = \frac{10!}{3!\,7!} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120$$

For combinations, we simply enter:

```
       3!10
120
```

What are the total number of poker hands? In this case we have 52 cards selected 5 at a time without regard to order:

```
       5!52
2598960
```

In summary, the five counting rules are

Identical Trials: n^k

Non-identical trials: $n_1 n_2 \cdots n_k$

Factorial: $n! = n \times (n-1) \times \cdots \times 3 \times 2 \times 1$

Permutations: ${}_nP_k = \dfrac{n!}{(n-k)!}$

Combinations: ${}_nC_k = \dbinom{n}{k} = \dfrac{n!}{k!(n-k)!}$

2.6 Random Variables and Probability Distributions

A **probability distribution** is set of possible values each having a certain probability, the sum of which totals one. Suppose we take a die and roll it. Since each outcome is equally likely, we call this distribution the `uniform` distribution. The probability distribution (relative frequency) looks like this:

```
       X←1 2 3 4 5 6
```

```
      ⎕←RelFreq←X,[.5]÷6
1 0.16667
2 0.16667
3 0.16667
4 0.16667
5 0.16667
6 0.16667
```

A **random variable** takes on any value of its distribution. Thus, we can simulate the roll of a single die:

```
      6 uniform randomVariable 1
3
```

The 6 on the left tells us that we are choosing a number between 1 and 6 randomly. The 1 on the right tells us we are rolling a single die. We can simulate rolling two dice as in a game of craps as follows:

```
     6 uniform randomVariable 2
2 5
```

In the game of Yahtzee, we roll 5 dice:

```
    6 uniform randomVariable 5
5 6 3 2 3
```

What is the expected value (average) roll of a single die?

$$E[X] = \sum xP(x)$$

Here we will simply multiply each value by its corresponding probability and add them up. For the uniform distribution:

```
      mean RelFreq
3.5
```

The variance of a random variable is defined as:

$$Var[X] = \sum (x - E[X])^2 P(x)$$

For the uniform distribution, the variance is:

```
      var RelFreq
2.9167
```

Of course the standard deviation $\sigma = \sqrt{Var[X]}$ is more meaningful because it is in the same units as the mean:

```
    sdev RelFreq
1.7078
```

Consider two stocks A and B valued at $10 each per share. Each stock is from a different industry and the performance of stock A is independent of Stock B. After one year, the stock prices will go up or down according to the following distribution:

		Stock A			**Stock B**
Event	**Value**	**Probability**	**Event**	**Value**	**Probability**
Up	**$15**	**60%**	**Up**	**$15**	**60%**
Down	**$ 5**	**40%**	**Down**	**$ 5**	**40%**

We have $20 available to invest. Should we buy two shares of Stock A or one share of stock A and one share of stock B? Does it matter?

We define the distribution as `multinomial` which contains the following parameters:

```
      PAYOUTS←15 5
      PROBS←.6 .4
     STOCK←PAYOUTS,[.5]PROBS
15 0.6
 5 0.4
```

To find probability of a $15 payout for Stock A, we simply apply the probability operator to the probability distribution STOCK:

```
       STOCK prob = 15
0.6
```

What is the expected payout of Stock A? We simply multiply each payout by its probability and add up the products. The `theoretical` operator applied to the `multinomial` distribution function will accomplish this:

```
      EA←mean STOCK
      EA
11
```

To find the variance of a random variable, we simply enter:

```
      VARA←var STOCK
      VARB←var STOCK
      VARA VARB
24 24
```

The standard deviation is a more useful measurement because it is in the same units as the expected value:

```
      sdev STOCK
4.89898
```

Which is riskier for the investor: owning two shares of stock A, or one share of Stock A and one share of Stock B?

The probability distribution for two shares of stock A is:

```
   2 shares of  Stock A
Event   Value    Probability
Up      $30       60%
Down    $10       40%
```

The expected value and variance of holding 2 shares of stock A are:

```
       STOCKA2←(2×PAYOUT),[.5]PROBS
       mean STOCKA2
22
       var STOCKA2
96
```

Note that while the expected value is twice that of the single stock, the variance is four times as high.

Assuming that the events are independent, the probability distribution for one share of stock A and one share of stock B is:

```
      One Share of A and one share of B

EventA  EventB      Value           Probability
------  ------      ----------      ----------------
Up       Up         15+15 = 30      0.6 x 0.6 = 0.36
Up       Down       15+5  = 20      0.6 x 0.4 = 0.24
Down     Up         5+15  = 20      0.6 x 0.4 = 0.24
Down     Down       5+5   = 10      0.4 x 0.4 = 0.16
```

There are three possible outcomes: $30, $20 and $10. We combine the probabilities for $X = 20$: $.24 + .24 = .48$

```
      PAYOUTAB←30 20 10
      PROBSAB←.36 .48 .16
      STOCKAB←PAYOUTAB,[.5]PROBSAB
```

The expected value and variance for holding one share of A and one share of B are:

```
      mean STOCKAB
22
      var STOCKAB
48
```

Observe that while expected values are the same, the variance of holding two different stocks is one-half that of owning two shares of one of the stocks. Thus we can sum the variances if the two events are independent:

```
      VARA+VARB
48
```

What is the probability that we lose money if we invest in two shares of Stock A?

```
       STOCK2 prob< 20
0.4
       STOCKAB prob < 20
0.16
```

Here we can see that the probability of losing money on buying 2 shares of Stock A is 40%, while the probability of losing money on buying one share each of A and B is only 16%.

Linear combinations of means and variances of random variables

The expected value a constant times a random variable is the constant times the expected value:

```
      mean 2 times #.SD.Height
137.55
      2 times mean #.SD.Height
137.55
```

The expected value of the sum of two random variables is the sum of the expected values:

```
      mean #.SD.(Height+Weight)
238.55
      (mean #.SD.Height)+(mean #.SD.Weight)
238.55
```

The variance of a constant times a random variable is the constant squared times the expected value:

```
      var #.SD.Height
20.03645092
```

```
      var #.SD.Weight
1778.712127
      var 2 times #.SD.Height
80.146
      4 times var #.SD.Height
80.146
```

The variance of the sum of two random variables is the sum of the variances if the two variables are independent. If the random variables are not independent, we have to add twice the covariance:

```
      var #.SD.(Height+Weight)
2047.2
       #.SD.Height cov #.SD.Weight
124.23
  (var #.SD.Height)+(var #.SD.Weight)+2 times (#.SD.Height cov #.SD.Weight)
2047.2
```

If two variables are independent, their covariance is zero.

The rules of expected value and variance are summarized below:

	Constant	**Constant Multiple**	**Sum of Two Random Variables**
Expected Value	$E[c]$=c	$E[aX] = aE[X]$	$E[X \pm Y] = E[X] \pm E[Y]$
Variance	$Var[c]$=0	$Var[aX] = a^2E[X]$	$Var[X \pm Y] = E[X] + E[Y] \pm 2\mathrm{cov}(X,Y)$

Bernoulli Trials

A Bernoulli trial has three characteristics:

1. Two possible outcomes success(1) and failure(0).
2. The probability of success (p) is the same on every trial
3. The trials are independent.

A simple example of a Bernoulli trial is a coin toss. There are two outcomes: heads (1) and tails (0); the probability of success (heads) is always ½, and each coin toss is independent of the others. To create a single Bernoulli trial with p=1/2, we use the `binomial` function without a left argument:

```
      binomial probability = 1
0.5
```

To simulate Bernoulli trials, we can use the `randomVariable` operator. Simulate 10 coin tosses:

```
      binomial randomVariable 10
1 1 0 1 1 0 1 0 0 1
```

Suppose the probability of success is not exactly ½. Let us define success as rolling a die and getting a six. Then p=1/6. We provide a left argument to the binomial function.

```
       P←1 div 6
       P binomial probability = 0 1
0.83333 0.16667
```

Let us simulate tossing 5 dice:

```
       P binomial randomVariable 5
0 0 0 1 0
```

The ones represent rolling a six; the zeros represent not rolling a six.

2.6.1 Discrete Distributions

A discrete distribution is defined by its probability mass function.

General Discrete Probability Distribution Formulas	
Characteristic	**General Formulas**
Expected Value (Mean)	$E(X) = \mu = \sum xP(x)$
Variance	$Var(X) = \sigma^2 = \sum (x-\mu)^2 P(x)$
Standard Deviation	$\sigma = \sqrt{\sum (x-\mu)^2 P(x)}$
Cumulative Probability	$P(X < x) = \sum_{i=0}^{x-1} P(X = i)$ $\quad P(X \geq x) = 1 - P(X < x)$ $P(X \leq x) = \sum_{i=0}^{x} P(X = i)$ $\quad P(X > x) = 1 - P(X \leq x)$

Discrete Distribution Characteristics and Formulas

Is zero a legitimate value?	**Upper bound**	**Distribution**	**Probability**	**Parameters**	**Mean**	**Variance**
Yes	Yes	Binomial*	$\binom{n}{x} p^x (1-p)^{n-x}$	n = sample size p = probability of success	np	$np(1-p)$
Yes	Yes	Hyper-Geometric**	$\frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$	m=successes in population n=sample size N=population size	$\frac{nm}{N}$	$\frac{nm(N-m)(N-n)}{N^2(N-1)}$
Yes	No	Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$	$\lambda = np$ Mean number of successes	λ	λ
No	Yes	Uniform	$\frac{1}{n}$	n = number of unique values	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
No	No	Geometric	$p(1-p)^{x-1}$	p = probability of success	$\frac{1}{p}$	$\frac{1-p}{p^2}$

*Infinite population or sampling with replacement

**Sampling without replacement from finite population

The Uniform Distribution

In the uniform distribution, each outcome is equally likely. An example of the uniform distribution is tossing a die. Since there are 6 possible outcomes, we provide the left argument *n* = 6. The uniform function calculates the probability of each outcome. For example the probability of rolling a two is:

```
    6 uniform 2
0.16667
```

In fact, the probabilities for any value between 1 and 6 are all 1/6:

```
      6 uniform 1 2 3 4 5 6
0.16667 0.16667 0.16667 0.16667 0.16667 0.16667
```

To calculate probabilities of events requires the use of the probability operator. Thus the probability of rolling a number greater than 4 is 1/3:

```
    6 uniform probability > 4
0.33333
```

The probability of not rolling a five is 5/6:

```
    6 uniform probability ne 5
0.83333
```

The Binomial Distribution

Suppose we flip 5 coins. What is the probability that exactly two of them turn up heads? For this we use the binomial distribution. In this case we have five Bernoulli trials. There are two parameters for the binomial distribution: the number of trials and the probability of success. Thus the binomial function produces the probability of the outcome.

```
     5 0.5 binomial 2
0.3125
```

If we want to find the probability of the event that at least two are heads, we must use the probability operator

```
     5 0.5 binomial probability ge 2
0.8125
```

A baseball player has a batting average of .221. Assuming that in 9-inning game a batter gets four at-bats. What is the probability that he bats safely (gets at least one hit)?

```
       4 .221 binomial probability ge 1
0.63174
```

At a car dealership, the probability that a customer will buy a car is 20%. If the dealership averages 15 customers a day, what is the expected number of sales? What is the variance?

```
     15 0.2 binomial theoretical mean 0
3
     15 0.2 binomial theoretical var 0
2.4
```

What is the probability that the dealership sells exactly 3 cars?

```
     15 0.2 binomial probability = 3
0.25014
```

What is the probability that the dealership sells at least 3 cars?

```
    15 0.2 binomial probability ge 3
0.60198
```

What is the probability that the dealership sells more than 2 cars but less than 6 cars?

```
    15 0.2 binomial probability between 2 6
0.54093
```

The Poisson Distribution

The `poisson` distribution is often used to model the number of random arrivals at a facility. Suppose that a hospital emergency room averages 5 patients per hour. What is the probability that in the next hour there will be exactly 6 emergencies?

```
   5 poisson probability = 6
0.14622
```

Suppose there are 10 seats in the waiting room. What is the probability that there will be enough seats for the patients who have to wait?

```
    5 poisson probability le 10
0.9863
```

Generate random arrivals at the emergency room for the next 24 hours:

```
    5 poisson randomVariable 24
3 3 9 5 4 4 11 5 1 4 7 7 6 3 7 5 5 4 4 4 8 2 3 5
```

The HyperGeometric Distribution

The `hyperGeometric` distribution is used to model sampling from a finite population. What is the probability of getting a pair of aces in a poker hand? In this case we have 4 aces (successes), 5 cards in the hand (sample) and 52 cards in the deck (population). The probability is about 4%:

```
    4 5 52 hyperGeometric probability = 2
0.03992981808
```

The probability of at least two aces is slightly greater:

```
    4 5 52 hyperGeometric probability ge 2
0.04168436605
```

2.6.2 Continuous Distributions

Continuous variables include quantities that can be infinitely subdivided, such as height, weight, cost and temperature. These variables cannot be represented by integers.

In Scotland, the trains from Dalmuir to Glasgow's Queen Street Station depart every 20 minutes. When you arrive at the station, how long do you have to wait for the next train? In this case we use the continuous version of the uniform distribution also known as the rectangular distribution. To illustrate, we will draw a graph and make the area under the curve equal to one. That way, probability equals area.

For the rectangular distribution, $a = 0$ and $b = 20$. The area under the curve equals 1, so the height $f(x) = 1/20$.

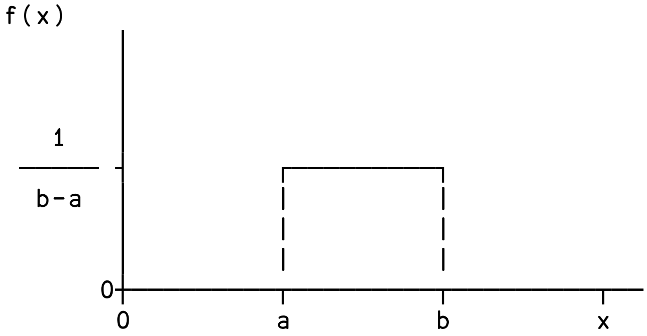


Figure 2 Rectangular (continuous uniform) distribution

What is the probability the next train will arrive within 10 minutes?

```
     20 rectangular probability<10
0.5
```

What is the probability that you will have at least 5 minutes to buy a ticket before the train arrives?

```
    20 rectangular probability ge 5
0.75
```

What is the probability it will take exactly 10 minutes?

```
     20 rectangular probability = 10
0
```

If this answer surprises you, it's because the probability that the train arrives in exactly 10 minutes is not the same as approximately 10 minutes. If we want to say that the train arrives in the 10th minute, we must indicate the probability as follows:

```
    20 rectangular probability between 9.5 10.5
0.05
```

This is true of all continuous distributions; we must talk about intervals, not exact values.

Exponential Distribution

Suppose the average life of a light bulb is 1000 hours. What is the probability that a light bulb will last 1200 hours? The poisson parameter lambda is 1/average life = 1/1000:

```
     .001 exponential probability > 1200
0.30119
```

The lifetimes of light bulbs do not follow a normal distribution. It is highly skewed to the right. This means that more than half of the light bulbs will fail before the average life of 1000 hours. In fact more than half of the light bulbs will fail before 700 hours:

```
    .001 exponential theoretical skewness 0
2
    .001 exponential theoretical median 0
693.15
```

The exponential distribution is related to the Poisson distribution. If we average 5 customers per hour in a store, the time between customer arrivals follows an exponential distribution with parameter λ= 5. Thus if a customer arrives at our store at exactly 10:00 a.m., what is the probability that the next customer will arrive within the next quarter hour?

```
    5 exponential probability < .25
0.7135
```

Normal Distribution

The normal distribution has a bell-shaped curve and is one of the most well-known distributions in statistics. It was discovered by Karl Friedrich Gauss. Below is a copy of the German 10 Mark note (before the Euro took over).

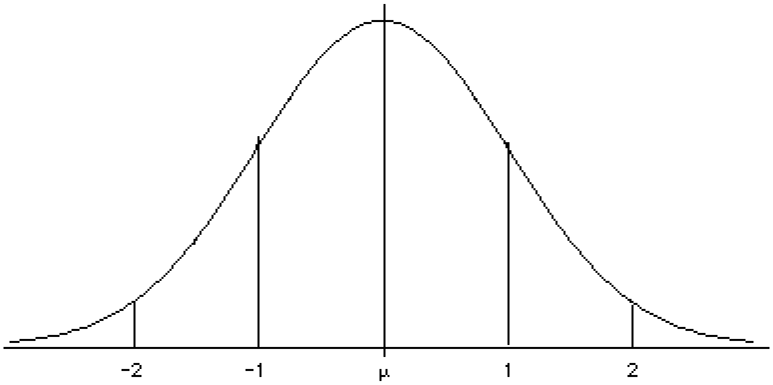
When Gauss was young, for misbehaving, he was given the arduous task of adding up the numbers from 1 to 100. He wrote the numbers from 1 to 50 in a column, and then wrote the numbers from 51 to 100 in the next column in reverse order. He then added the rows: $1 + 100 = 101$; $2 + 99 = 101$, etc. and discovered there were 50 rows each containing the sum of 101. The solution was simple: $50 \times 101 = 5050$.

The formula for the density function of the normal distribution is: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, where μ is the mean and σ is the standard deviation. If we set $\mu = 0$ and $\sigma = 1$ we have the standard normal distribution which corresponds to Z-scores that we discussed previously; the formula simplifies to:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

The bell-shaped curve can be seen in the density function:

```
      normal ¯3 ¯2 ¯1 0 1 2 3
0.0044318 0.053991 0.24197 0.39894 0.24197 0.053991 0.0044318
```



As we have seen with other continuous distributions, particularly with the rectangular, the density function does not give us probabilities; we need to calculate the area under the curve. The density function for the standard normal distribution is impossible to integrate symbolically; we must integrate numerically. Many textbooks provide tables for the standard normal distribution; TamStat allows us to simply enter the function monadically without parameters:

```
     normal probability < 1.25
0.89435
```

We can also find the value of the standard normal distribution which cuts off the top 5% of values:

```
     normal criticalValue  < .05
1.6449
```

What is the 30th percentile of the standard normal distribution?

```
     normal criticalValue  > .30
¯0.5244
```

In the real world, we must translate values so that they match the values in the standard normal distribution. Suppose male heights are normally distributed with a mean of 68 inches and a standard deviation of 3 inches. What is the probability that a randomly selected man is no taller than 72 inches?

```
     68 3 normal probability < 72
0.90879
```

What is the probability that a randomly selected man is between 60 and 70 inches?

```
     68 3 normal probability between 60 70
0.74368
```

What height represents the 95th percentile of all male heights?

```
     68 3 normal criticalValue > .95
72.935
```

Normal Approximation to Binomial Distribution

A Boeing 747 seats 450 passengers. Assuming 5% no-show rate, an airline overbooks this flight by 14 passengers. What is the probability that the airline has to turn away ticketed passengers?

We will assume a binomial distribution with $n = 450 + 14 = 464$, and $p = 1 - 0.95 = 0.05$. x is the number of passengers showing up for the flight.

```
      464 .95 binomial probability > 450
0.013939
```

We can also use the normal approximation to the binomial when both $np > 5$ and $n(1 - p) > 5$.

```
     .05 .95 x 464
23.2 440.8
```

We set $\mu = np$ and $= \sqrt{np(1 - p)}$. Thus

```
      MU←464 times .95
440.8
      SIGMA←sqrt 464 times 0.95 times 1-0.95
4.6947
```

We also must use the continuity correction factor for the normal distribution. $X = 450$ for the binomial distribution translates to $449.5 \leq X < 450.5$ in the normal distribution; thus $X > 450$ in the binomial would be $X > 450.5$ in the normal distribution.

```
      MU SIGMA normal probability > 450.5
0.019406
```

We see that the airline will turn away ticketed passengers less than 2% of the time.

Triangular Distribution

When not much is known about the shape of a distribution, we can use the triangular distribution when only the minimum, maximum and most likely values are known. Suppose a construction project usually takes 5 days, but can be completed in as few as four days under favorable conditions, but may take as many as 7 days under unfavorable conditions. What is the probability that the project can be completed in less than six days?

```
      4 5 7 triangular probability < 6
0.8333333333
```

Continuous Distribution Characteristics and Formulas

Lower bound	Upper bound	Mode	Distribution	Density	Parameters	Mean	Variance
Yes	Yes	No	Rectangular	$\frac{1}{b-a}$	a = lower bound b = upper bound	$\frac{\boldsymbol{a+b}}{\mathbf{2}}$	$\frac{(b-a)^2}{12}$
Yes	Yes	Yes	Triangular	For $x \leq c$: $\frac{2(x-a)}{(b-a)(c-a)}$ For $x > c$: $\frac{2(b-x)}{(b-a)(b-c)}$	a=lower bound c=mode b=upper bound	$\frac{a+b+c}{3}$	$\frac{a^2+b^2+b^2}{18} - \frac{ab+ac+bc}{18}$
Yes (0)	No	Yes	Exponential	$\lambda e^{-\lambda x}$	λ = Average rate of arrivals, defects, successes, etc.	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
No $(\mu - 3\sigma > 0)$	No	Yes	Normal	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ = mean σ = standard deviation	μ	σ^2

Notes:

1. Rectangular distribution is also known as the continuous uniform distribution.
2. Triangular distribution is used when nothing else is known about the shape of the distribution
3. Exponential distribution is the continuous equivalent of the geometric distribution. It also measures the time between arrivals in the Poisson distribution
4. Normal distribution can also be used as an approximation to the binomial distribution.

2.6.2 Scalar Characteristics of Distribution Functions

Distribution functions perform as scalar functions in most cases. If the right argument is an array, the result is an array with the same shape:

```
    normal 0 1 2                    ⍝ Density
0.39894 0.24197 0.053991
    normal prob < 0 1 2             ⍝ Cumulative probability
0.5 0.84134 0.97725
    normal critVal < .1 .05 .01     ⍝ Inverse
1.2816 1.6449 2.3263
```

For single-parameter distributions, either the left and right arguments must be scalar, or both must have the same shape:

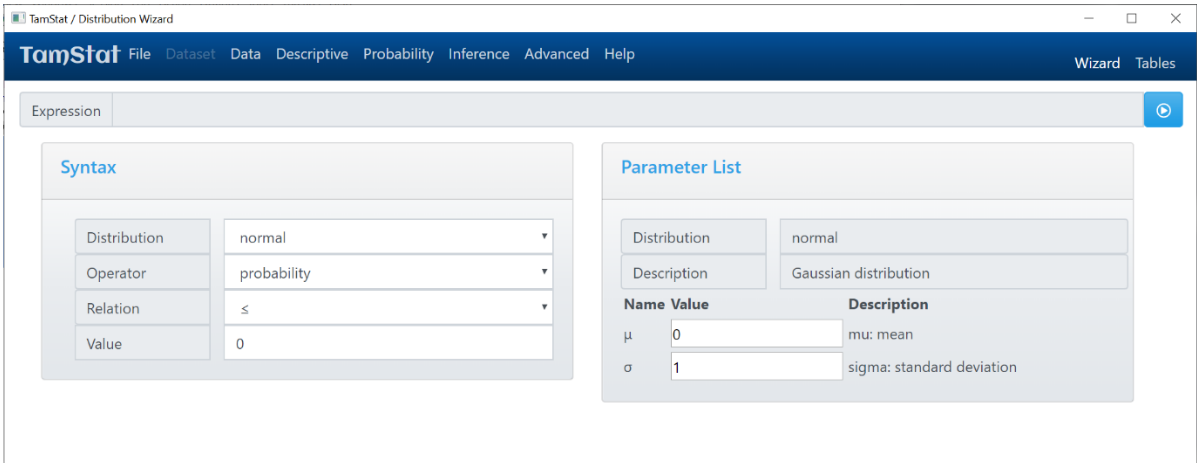
```
     4 poisson 2              ⍝ P(X=2|λ=4)
0.14653
     4 poisson 1 2 3          ⍝ P(X=1|λ=4);P(X=2|λ=4);P(X=3|λ=4)
0.073263 0.14653 0.19537
     1 2 3 poisson 4          ⍝ P(X=4|λ=1);P(X=4|λ=2);P(X=4|λ=3)
0.015328 0.090224 0.16803
     1 2 3 poisson 4 5 6      ⍝ P(X=4|λ=1);P(X=5|λ=2);P(X=6|λ=3)
0.015328 0.036089 0.050409
```

For multiple-parameter distributions, the left argument may be nested; the right argument and each item in the left argument must either be scalar or have the same shape:

```
     2 27 fDist critVal < .05              ⍝ All values are scalar
3.3541
     2 27 fDist critVal < .1 .05 .01   ⍝ Scalar parameters; vector right arg
2.5106 3.3541 5.4881
     (2 2 4)27 fDist critVal < .05        ⍝ 1st parameter is a vector
3.3541 3.3541 2.7278
     (2 2 4)27 fDist prob > 2.22 8.13 3.56 ⍝ 1st parm, right arg same shape
0.12806 0.0017228 0.018601
     (2 2 4)(27 4 27) fDist critVal < .05  ⍝ Both parameters same shape
3.3541 6.9443 2.7278
     (2 2 4)(27 4 27) fDist prob > 2.22 8.13 3.56 ⍝Parameters, right arg agree
0.12806 0.03898 0.018601
```

2.6.3 Distribution Wizard

Select “Probability” from the main menu then select “Distribution Wizard”. [1] The following screen will appear:



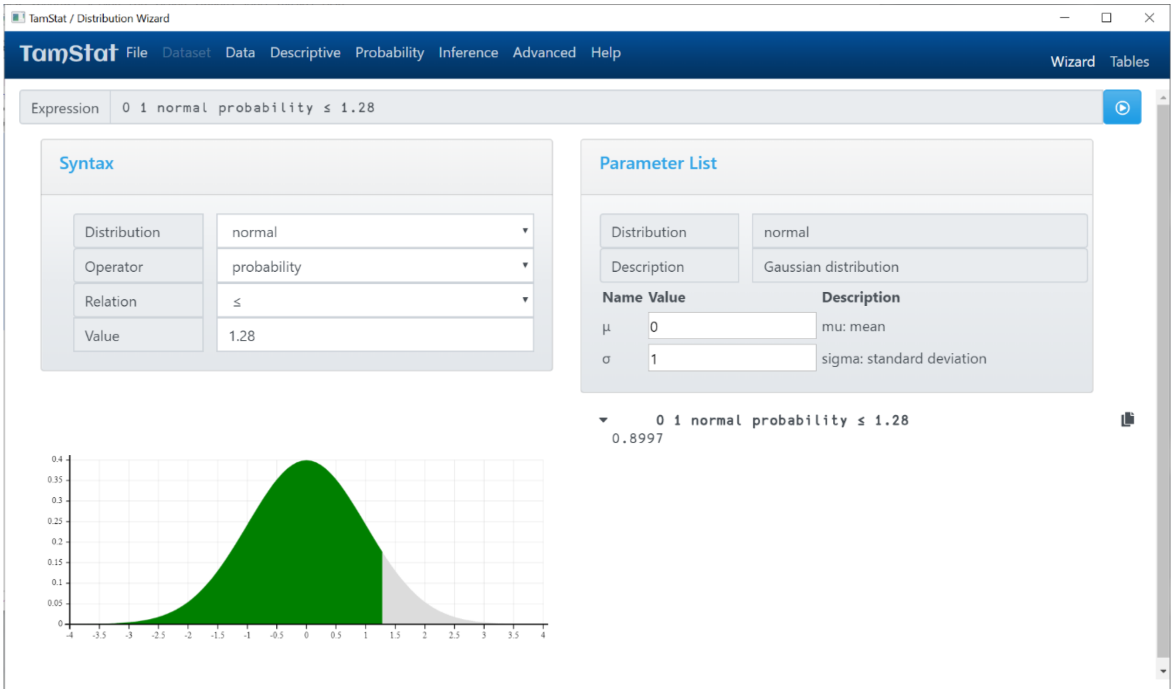
Distribution	normal
Operator	probability
Relation	≤
Value	0

Distribution	normal
Description	Gaussian distribution

Name	Value	Description
μ	0	mu: mean
σ	1	sigma: standard deviation

The syntax group contains the basic inputs required for the various distribution operations. Once a distribution is selected, the appropriate parameter list will appear in a grid on the right. The user can then enter the parameter values in the ‘Value’ column of the parameter grid.

Select the appropriate distribution, operator and relation. Then enter the distribution parameters in the parameter list on the right, and value of interest. Press “Calc” to generate the expression and display the result.



Expression `0 1 normal probability ≤ 1.28`

Distribution	normal
Operator	probability
Relation	≤
Value	1.28

Distribution	normal
Description	Gaussian distribution

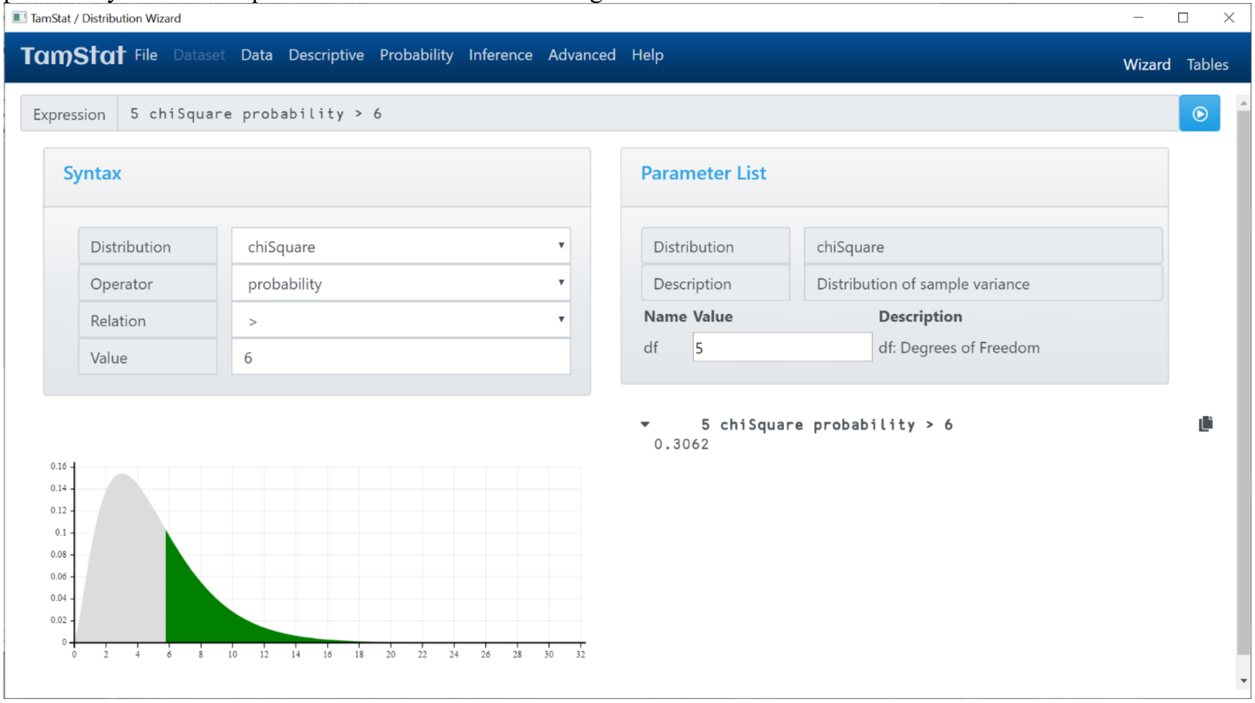
Name	Value	Description
μ	0	mu: mean
σ	1	sigma: standard deviation

```
      0 1 normal probability ≤ 1.28
0.8997
```

Press the “Graph” button to display the probability curve and area of interest when the “probability” or criiticalValue” operator is selected. Drag the thumb along the trackbar at the bottom of the graph to see how the cumulative probability changes with the value.

[1] From the Dyalog APL session, simply enter `buildDist ''`.

When performing a one-tail hypothesis test, one often wants to determine the p-value which is often an upper-tail probability. To obtain an upper tail probability, change the relation to greater-than (>). For example, find the probability that a chi-square random variable with 5 degrees of freedom exceeds the value 6:



Expression `5 chiSquare probability > 6`

Syntax

Distribution	chiSquare
Operator	probability
Relation	>
Value	6

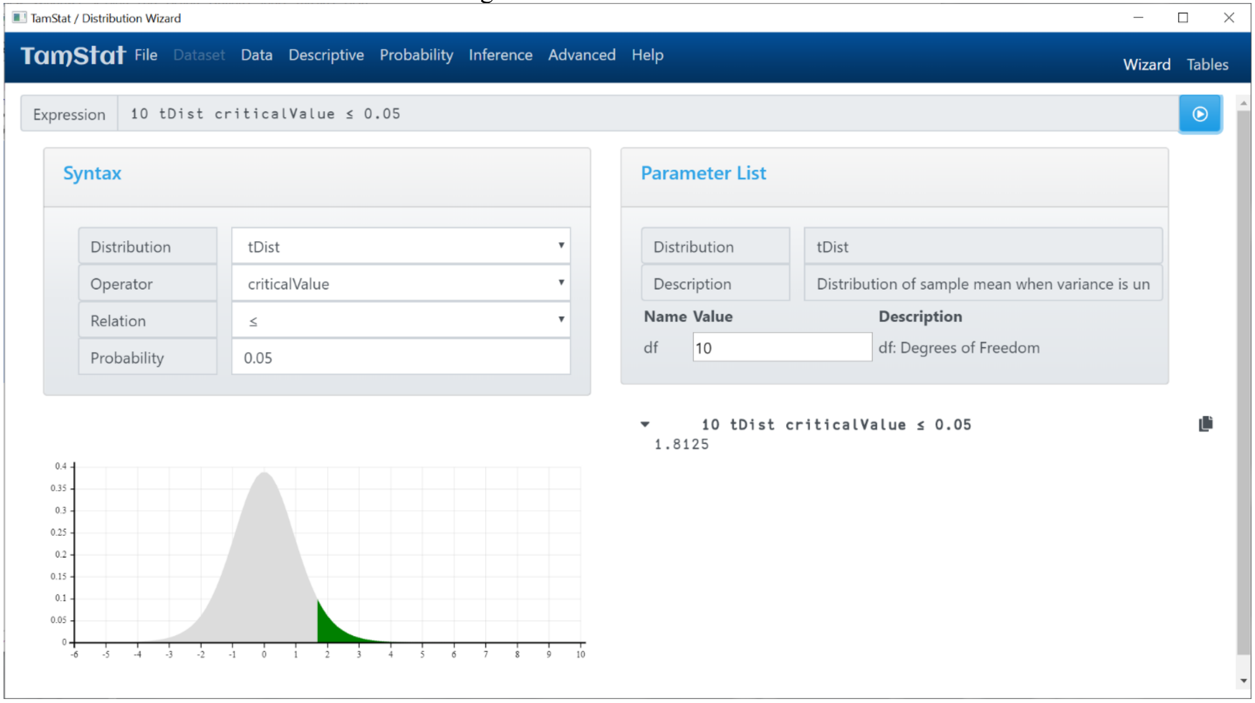
Parameter List

Distribution	chiSquare
Description	Distribution of sample variance

Name	Value	Description
df	5	df: Degrees of Freedom

```
    5 chiSquare probability > 6
0.3062
```

Sometimes one is more interested in the critical value than the probability. We are interested in finding the critical value of the Student t distribution with 10 degrees of freedom which is less than 5% of all values:



Expression `10 tDist criticalValue ≤ 0.05`

Syntax

Distribution	tDist
Operator	criticalValue
Relation	≤
Probability	0.05

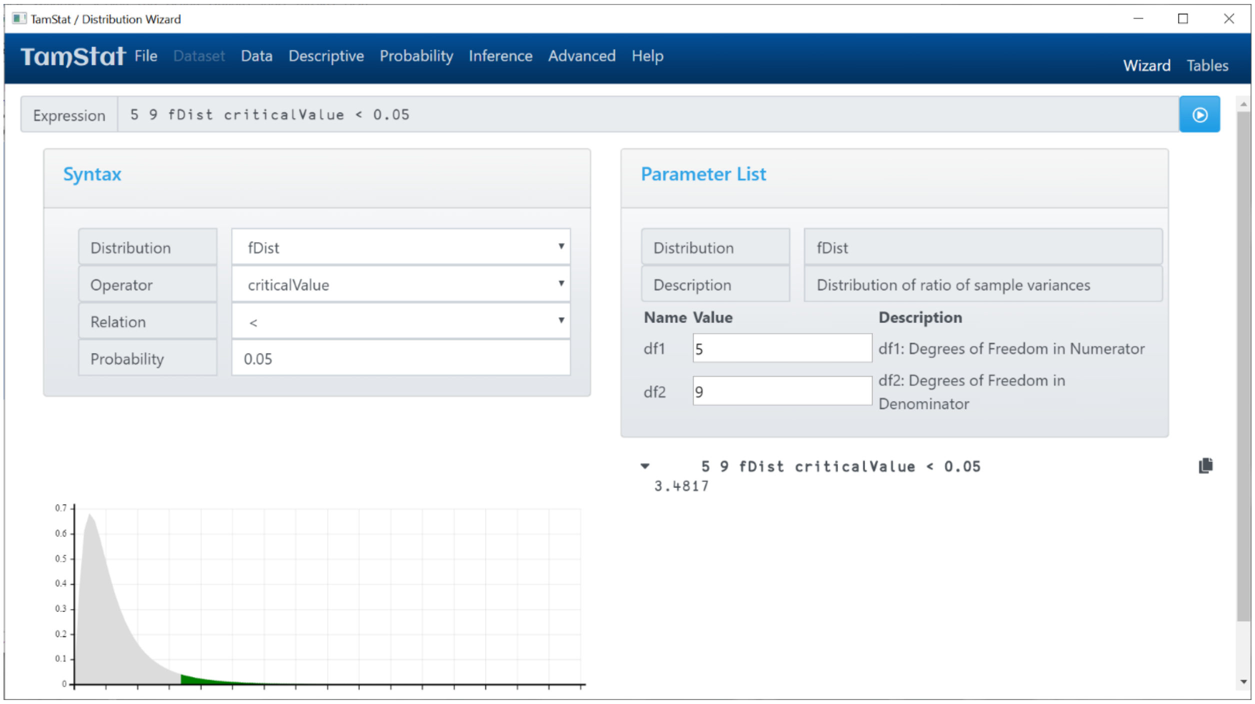
Parameter List

Distribution	tDist
Description	Distribution of sample mean when variance is un

Name	Value	Description
df	10	df: Degrees of Freedom

```
    10 tDist criticalValue ≤ 0.05
1.8125
```

When performing analysis of variance, one is interested in finding the upper-tail critical value for the F distribution:



Expression `5 9 fDist criticalValue < 0.05`

Syntax

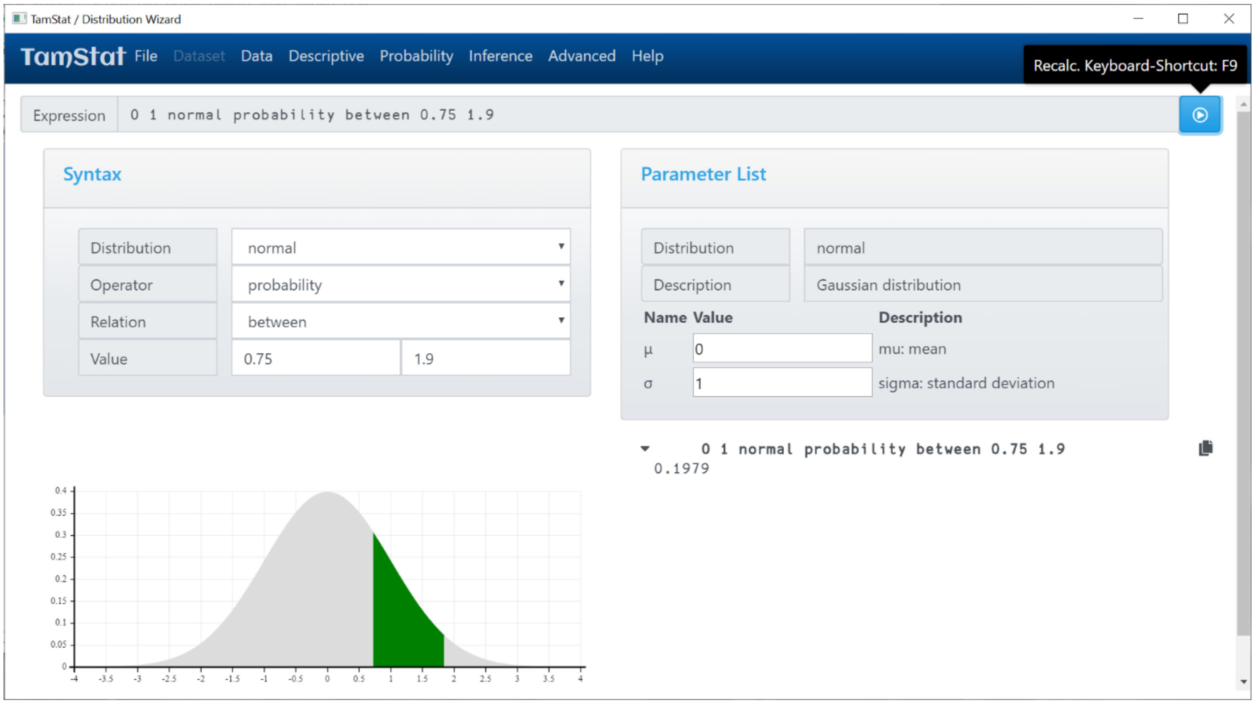
Distribution	fDist
Operator	criticalValue
Relation	<
Probability	0.05

Parameter List

Distribution	fDist	
Description	Distribution of ratio of sample variances	
Name	**Value**	**Description**
df1	5	df1: Degrees of Freedom in Numerator
df2	9	df2: Degrees of Freedom in Denominator

```
    5 9 fDist criticalValue < 0.05
3.4817
```

To determine the probability between two values, select the relation "between" and enter the two boundary values, "From" and "To":



Expression `0 1 normal probability between 0.75 1.9`

Syntax

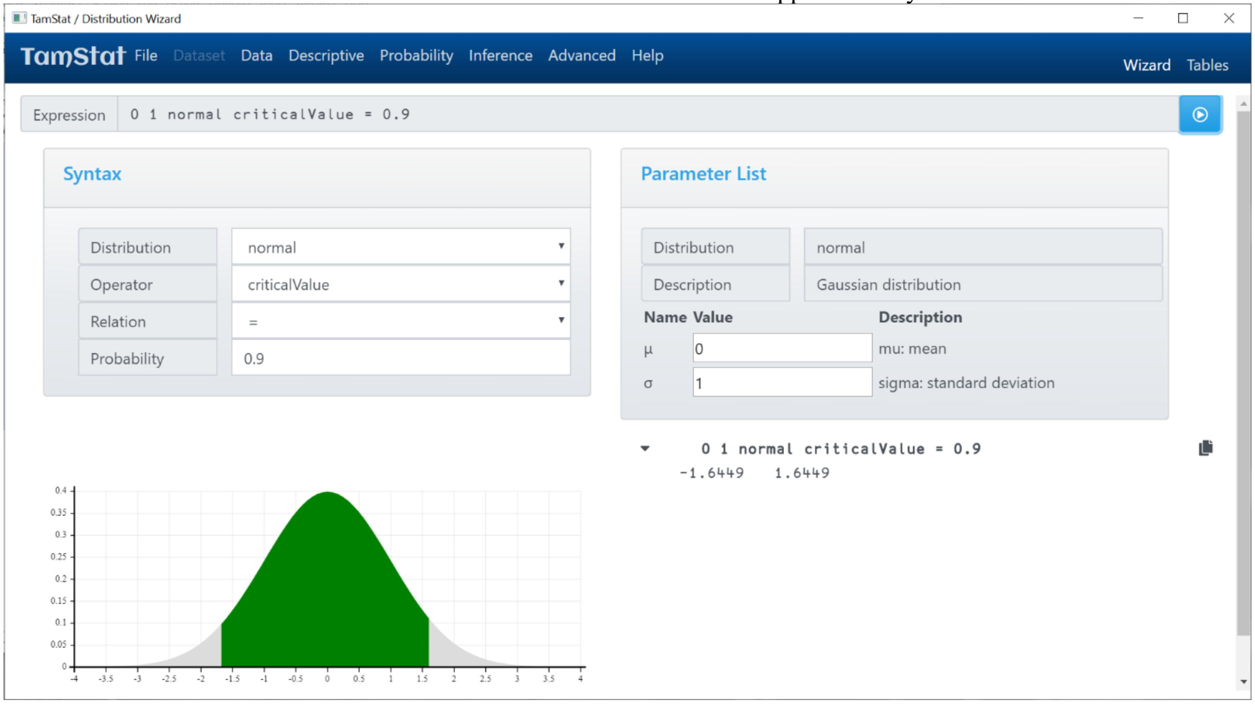
Distribution	normal	
Operator	probability	
Relation	between	
Value	0.75	1.9

Parameter List

Distribution	normal	
Description	Gaussian distribution	
Name	**Value**	**Description**
μ	0	mu: mean
σ	1	sigma: standard deviation

```
    0 1 normal probability between 0.75 1.9
0.1979
```

To find the critical values for a confidence interval, use the relation "=" or "outside" with the confidence level. Thus the critical value for the normal distribution at 90% confidence is approximately 1.645.



Expression: `0 1 normal criticalValue = 0.9`

Syntax

Distribution	normal
Operator	criticalValue
Relation	=
Probability	0.9

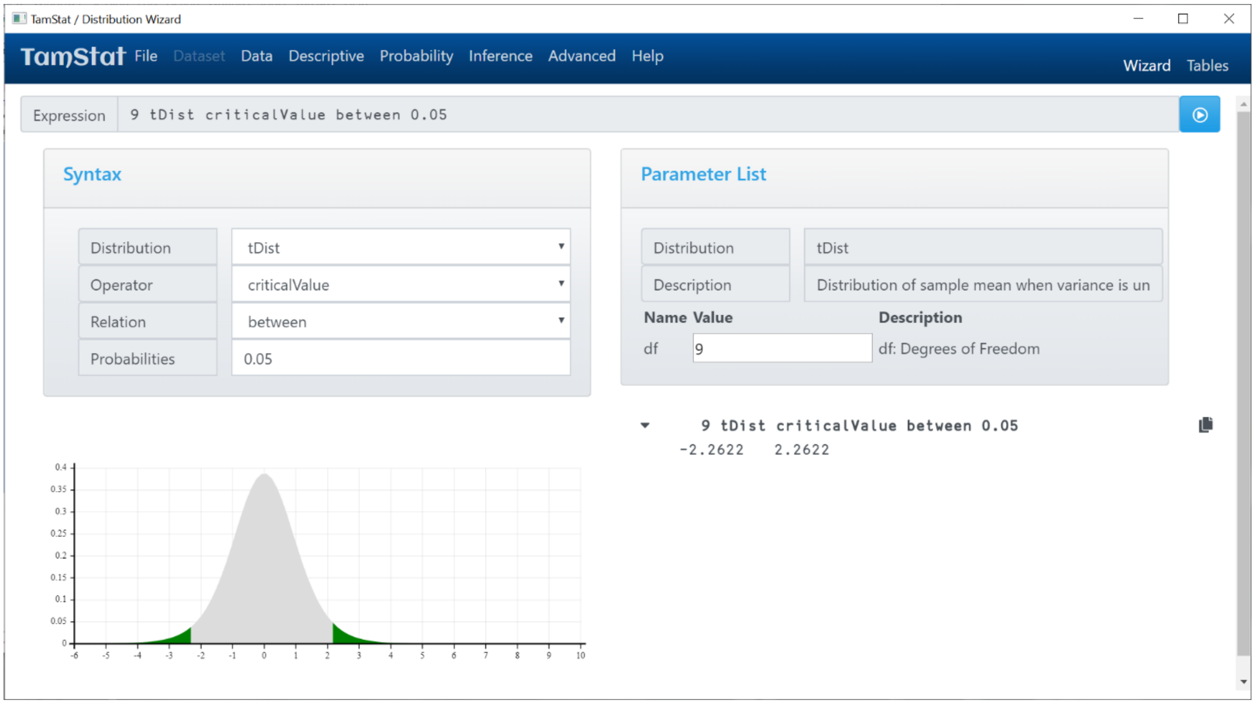
Parameter List

Distribution	normal
Description	Gaussian distribution

Name	Value	Description
μ	0	mu: mean
σ	1	sigma: standard deviation

```
    0 1 normal criticalValue = 0.9
 -1.6449    1.6449
```

To find the critical value for a two-sided hypothesis test, use the alternate hypothesis "≠" or "between" with the significance level α. Thus the critical value for a t-distribution with 9 degrees of freedom at distribution at α=0.05 is aproximately 2.2622:



Expression: `9 tDist criticalValue between 0.05`

Syntax

Distribution	tDist
Operator	criticalValue
Relation	between
Probabilities	0.05

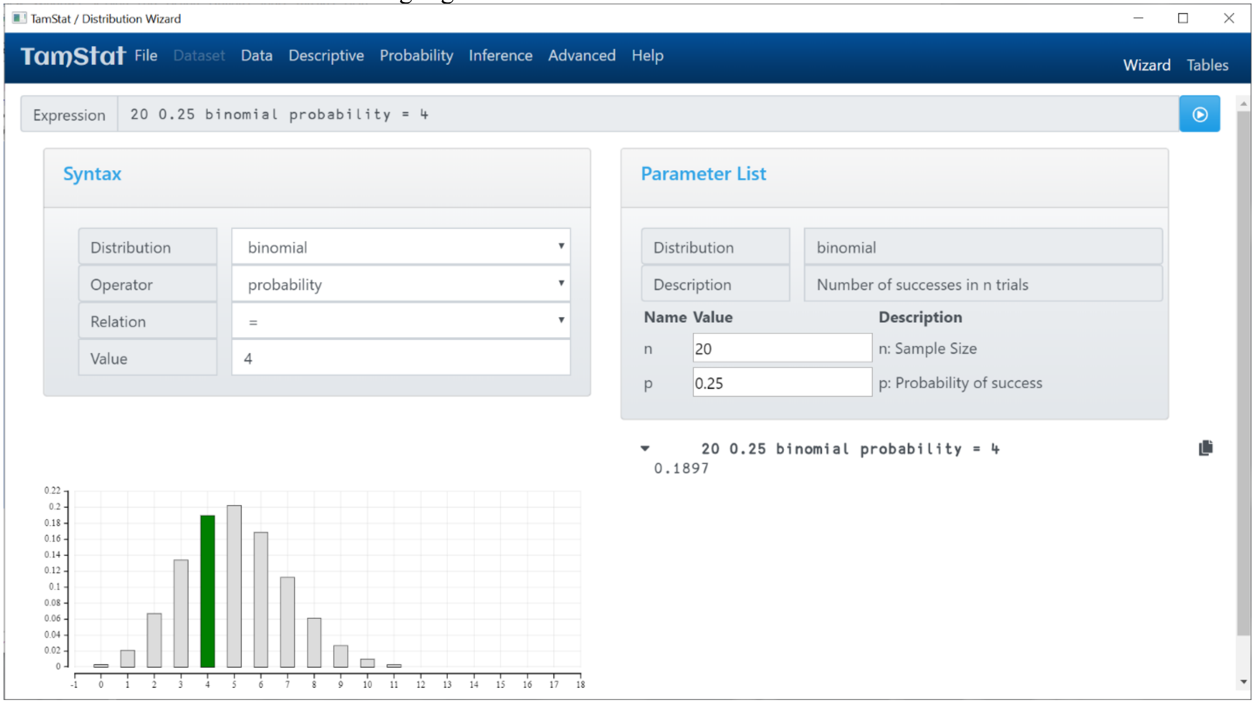
Parameter List

Distribution	tDist
Description	Distribution of sample mean when variance is un

Name	Value	Description
df	9	df: Degrees of Freedom

```
    9 tDist criticalValue between 0.05
 -2.2622    2.2622
```

When a discrete distribution is selected, a bar chart showing the probabilities for each individual value is displayed, with the value or values of interest highlighted.



Expression `20 0.25 binomial probability = 4`

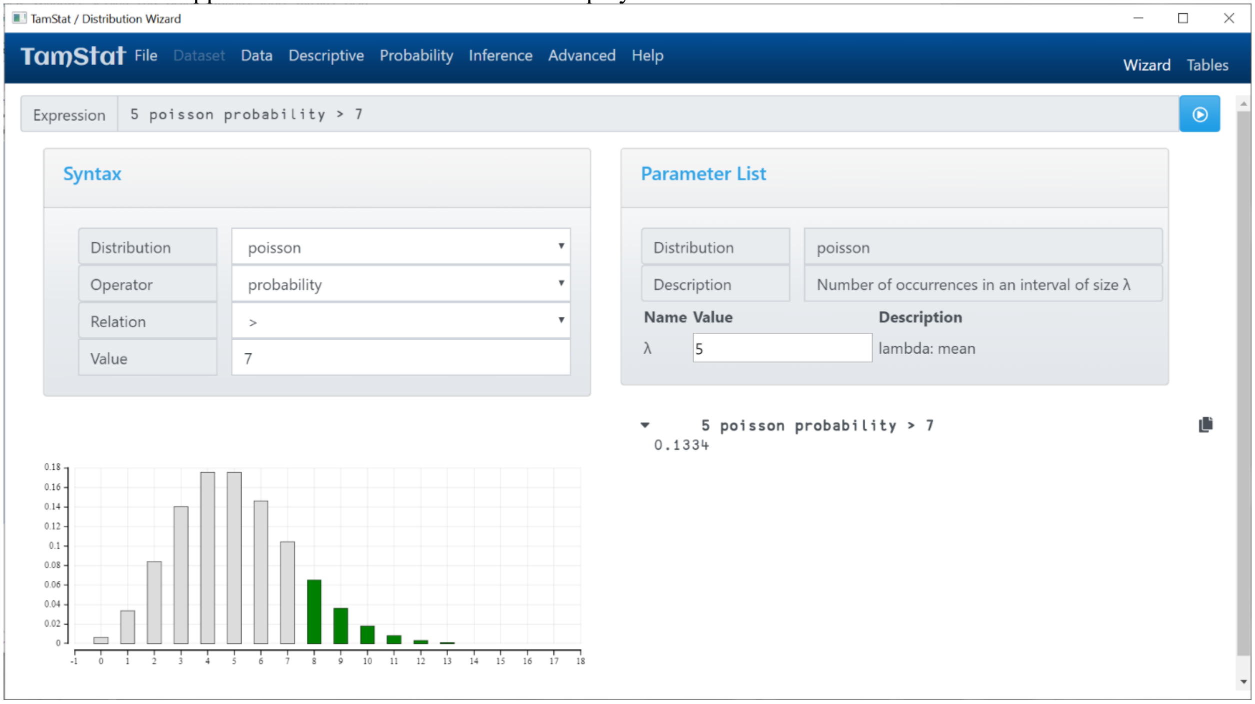
Distribution	binomial
Operator	probability
Relation	=
Value	4

Distribution	binomial
Description	Number of successes in n trials

Name	Value	Description
n	20	n: Sample Size
p	0.25	p: Probability of success

```
20 0.25 binomial probability = 4
0.1897
```

Cumulative and upper-tail distributions can also be displayed for discrete distributions:



Expression `5 poisson probability > 7`

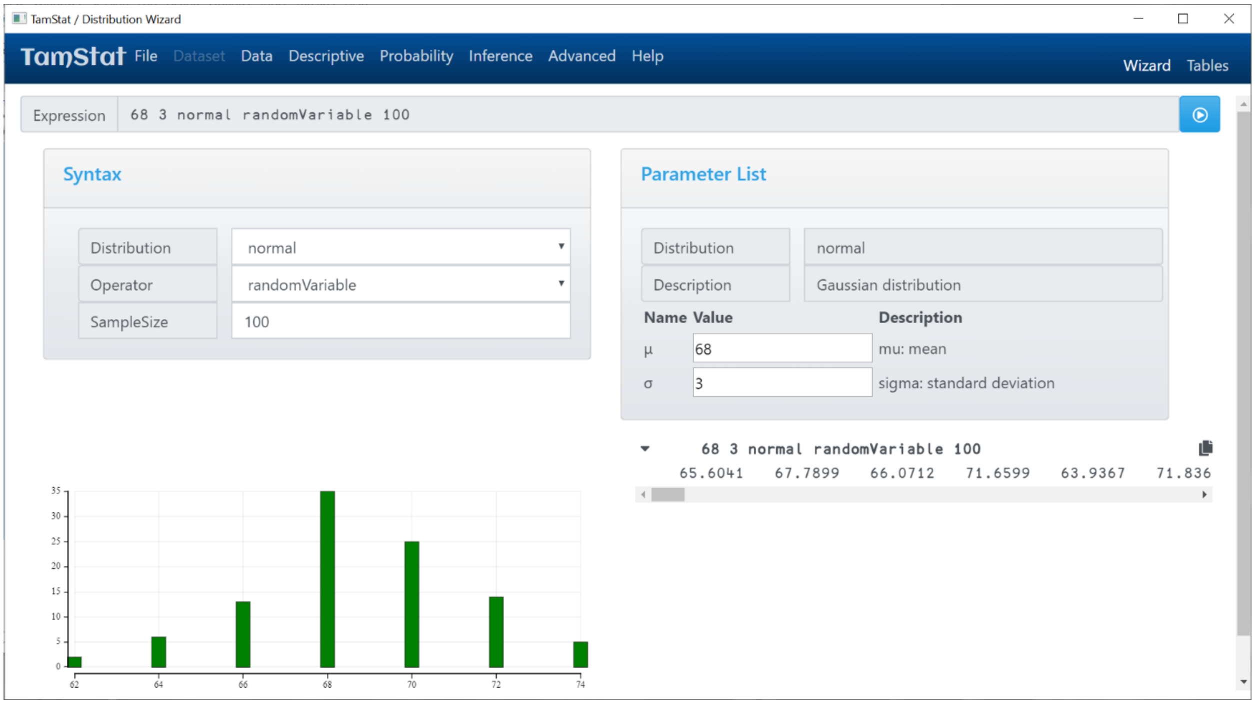
Distribution	poisson
Operator	probability
Relation	>
Value	7

Distribution	poisson
Description	Number of occurrences in an interval of size λ

Name	Value	Description
λ	5	lambda: mean

```
5 poisson probability > 7
0.1334
```

When you select `randomVariable`, the distribution wizard will generate a sample of size n from that distribution and display a histogram of the data generated:



Expression: `68 3 normal randomVariable 100`

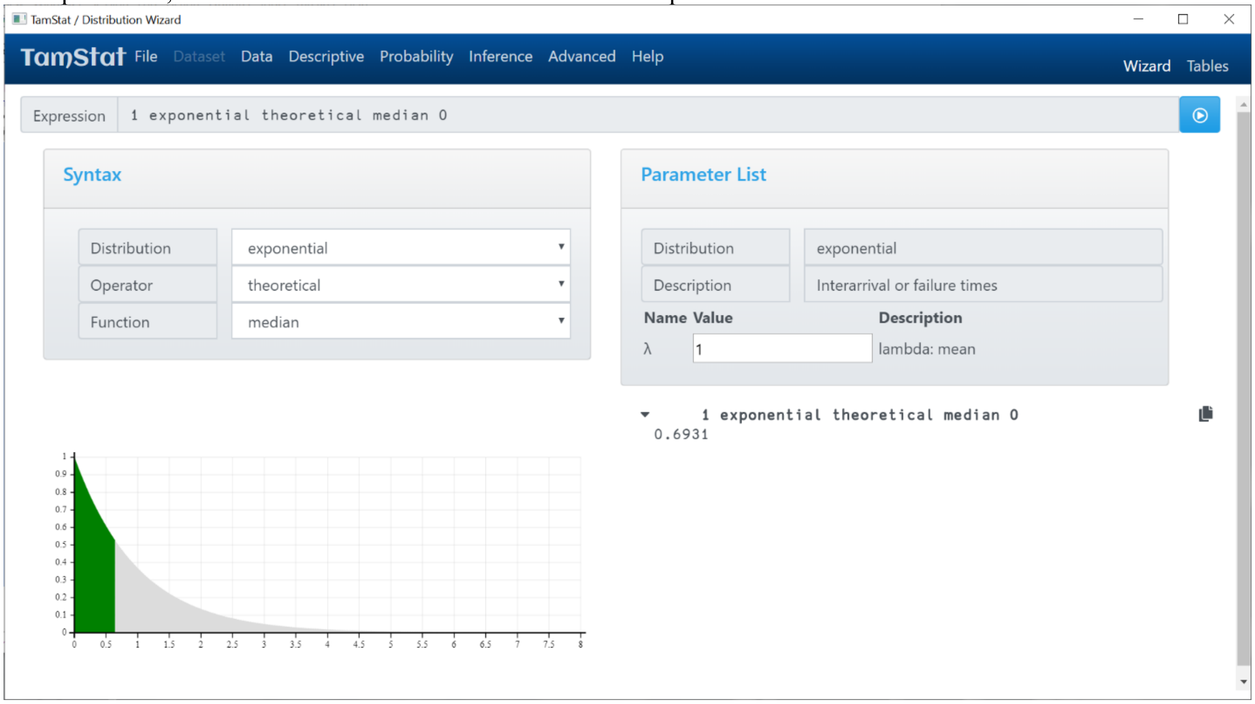
Distribution	normal
Operator	randomVariable
SampleSize	100

Distribution	normal
Description	Gaussian distribution

Name	Value	Description
μ	68	mu: mean
σ	3	sigma: standard deviation

```
      68 3 normal randomVariable 100
    65.6041    67.7899    66.0712    71.6599    63.9367    71.836
```

When theoretical is selected the appropriate parameter is displayed along with the density curve. In the example below, the median divides the distribution into two equal areas.



Expression: `1 exponential theoretical median 0`

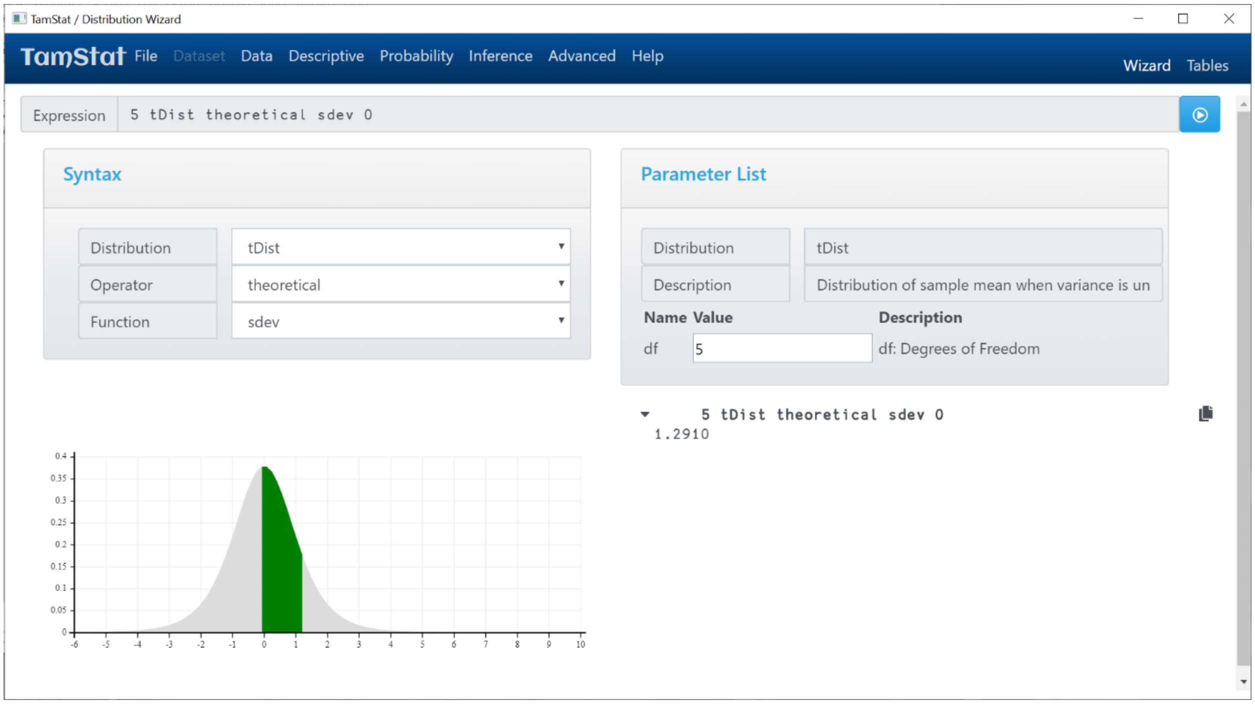
Distribution	exponential
Operator	theoretical
Function	median

Distribution	exponential
Description	Interarrival or failure times

Name	Value	Description
λ	1	lambda: mean

```
      1 exponential theoretical median 0
  0.6931
```

In another example, the standard deviation is represented by the area between the mean and one standard deviation above the mean. The standard deviation of the t-Distrtibution with 5 degrees of freedom is 1.29, the area between the mean (zero) and the standard deviation (1.29) is displayed:



Expression `5 tDist theoretical sdev 0`

Syntax

Distribution	tDist
Operator	theoretical
Function	sdev

Parameter List

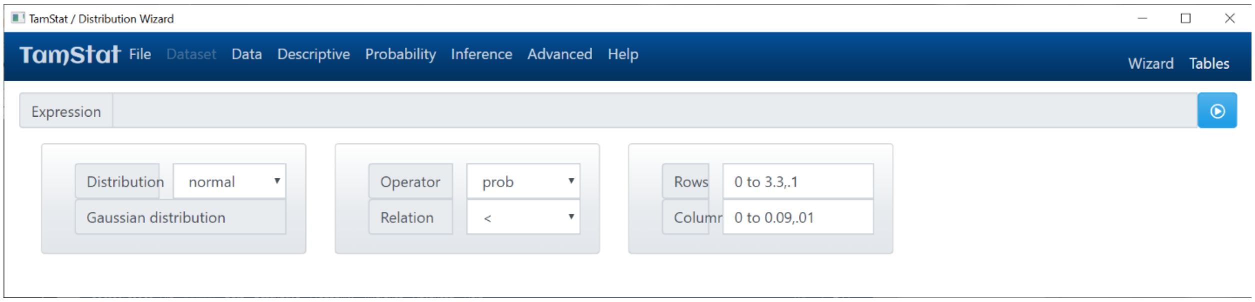
Distribution	tDist
Description	Distribution of sample mean when variance is un

Name	Value	Description
df	5	df: Degrees of Freedom

```
    5 tDist theoretical sdev 0
1.2910
```

2.6.4 Statistical Tables

To display a table of values from a statistical distribution, select “Propbability” from the main menu, then select “Distribution Tables”.   The following screen will appear,



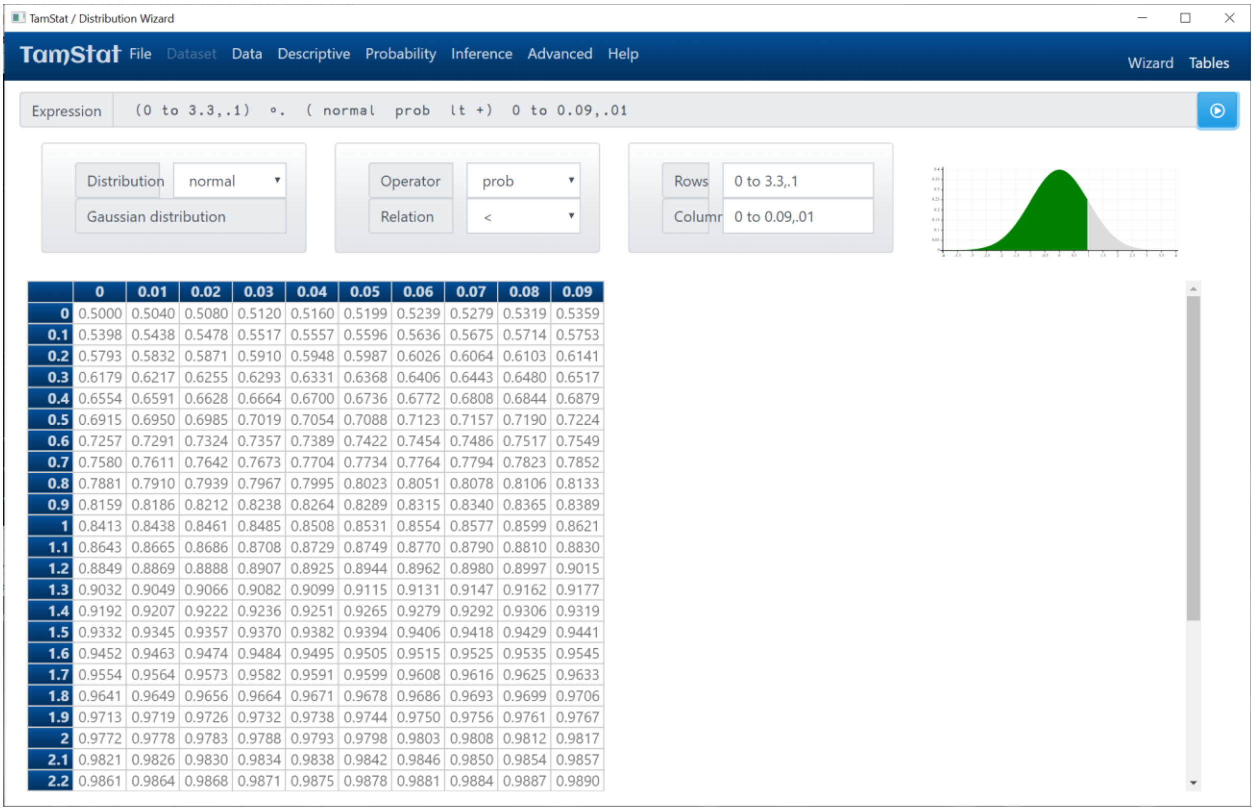
Expression

Distribution	normal
Gaussian distribution	

Operator	prob
Relation	<

Rows	0 to 3.3,.1
Columr	0 to 0.09,.01

Select “Calc”. A normal table will appear with a graphic depiction of the bell curve with the appropriate area shaded . Scroll bars appear where appropriate, and the window can be resized:



```
Expression   (0 to 3.3,.1) ∘. ( normal prob lt +) 0 to 0.09,.01
```

Distribution: normal
Gaussian distribution

Operator: prob
Relation: <

Rows: 0 to 3.3,.1
Columr: 0 to 0.09,.01

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890

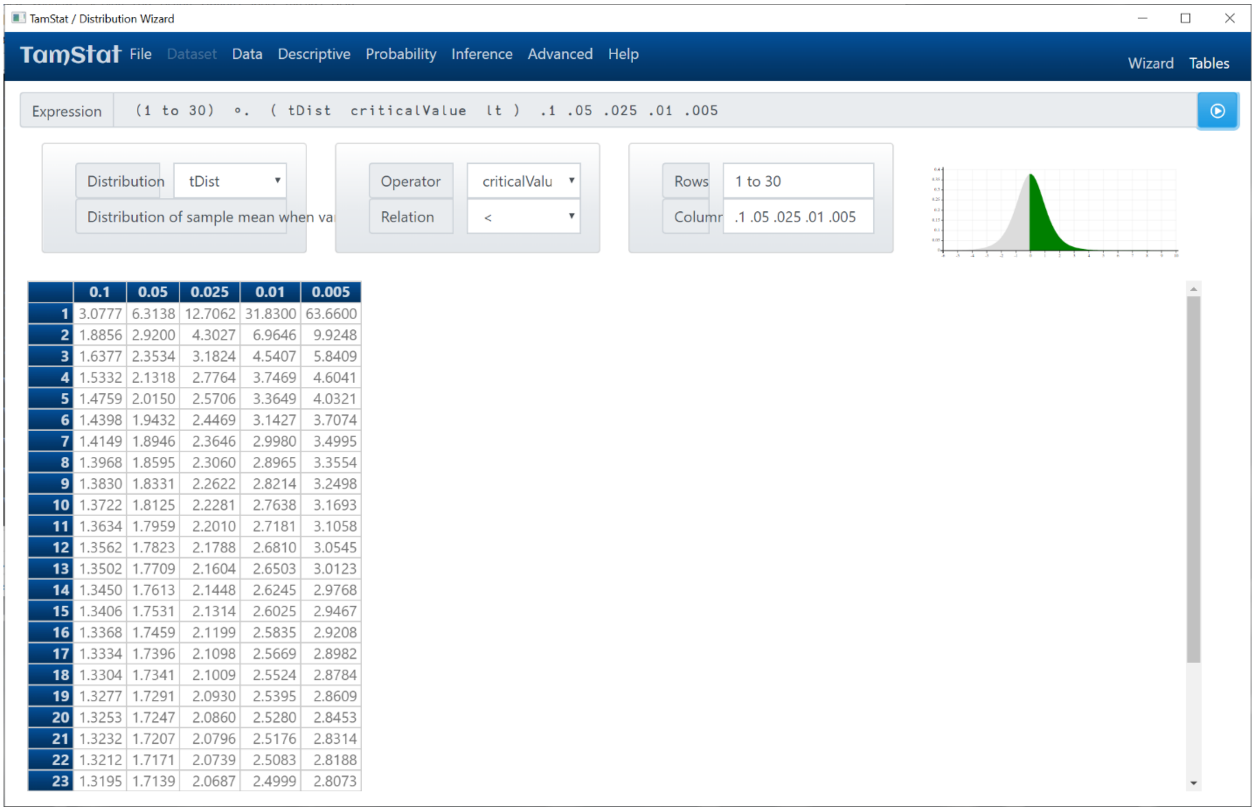
To obtain a different table, select the appropriate distribution from the drop-down. The remaining fields will be set to the appropriate defaults. For example, the t-distribution defaults to the **`criticalValue`** operator, where the rows are degrees of freedom and columns are upper tail probabilities. The rows and columns are simply lists of values used as inputs to the distribution. To generate a long sequence of values, use the dyadic **`to`** function. The left argument is the starting point and the right argument is the ending point. e.g:

```
      1 to 10
1 2 3 4 5 6 7 8 9 10
```

Note the step size defaults to 1. If you want to change the step size, say to one-tenth, append that value to the right argument:

```
      0 to 1,0.1
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
```

The following is an example of a one-tail probability obtain a different table, select the appropriate distribution from the drop-down. The remaining fields will be set to the appropriate



Expression `(1 to 30) ∘. ( tDist criticalValue lt ) .1 .05 .025 .01 .005`

Distribution: tDist

Distribution of sample mean when va

Operator: criticalValu

Relation: <

Rows: 1 to 30

Columr: .1 .05 .025 .01 .005

	0.1	0.05	0.025	0.01	0.005
1	3.0777	6.3138	12.7062	31.8300	63.6600
2	1.8856	2.9200	4.3027	6.9646	9.9248
3	1.6377	2.3534	3.1824	4.5407	5.8409
4	1.5332	2.1318	2.7764	3.7469	4.6041
5	1.4759	2.0150	2.5706	3.3649	4.0321
6	1.4398	1.9432	2.4469	3.1427	3.7074
7	1.4149	1.8946	2.3646	2.9980	3.4995
8	1.3968	1.8595	2.3060	2.8965	3.3554
9	1.3830	1.8331	2.2622	2.8214	3.2498
10	1.3722	1.8125	2.2281	2.7638	3.1693
11	1.3634	1.7959	2.2010	2.7181	3.1058
12	1.3562	1.7823	2.1788	2.6810	3.0545
13	1.3502	1.7709	2.1604	2.6503	3.0123
14	1.3450	1.7613	2.1448	2.6245	2.9768
15	1.3406	1.7531	2.1314	2.6025	2.9467
16	1.3368	1.7459	2.1199	2.5835	2.9208
17	1.3334	1.7396	2.1098	2.5669	2.8982
18	1.3304	1.7341	2.1009	2.5524	2.8784
19	1.3277	1.7291	2.0930	2.5395	2.8609
20	1.3253	1.7247	2.0860	2.5280	2.8453
21	1.3232	1.7207	2.0796	2.5176	2.8314
22	1.3212	1.7171	2.0739	2.5083	2.8188
23	1.3195	1.7139	2.0687	2.4999	2.8073

2.7 Exercises

1. Given that z is a standard normal random variable, computer the following probabilities:
   a. P(Z > 0.44)
   b. P(Z < 1.20)
   c. P(0 < Z <.83)
   d. P(.52<z<1.22)
2. Given that z is a standard normal random variable, find z for each situation:
   a. The area to the left of z is .9750
   b. The area to the right of z is .1314
   c. The area of the left of z is .6700
3. The average stock price for companies making up the S&P 500 is $30 and the standard deviation is $8.20. Assume the stock prices are normally distributed.
   a. What is the probability a company will have a stock price of at least $40?
   b. What is the probability a company will have a stock price higher than $20?
   c. How high does a stock price have to be to put a company in the top 10%?
4. An average of 15 aircraft accidents occur each year.
   a. Compute the mean number of aircraft accidents per month
   b. Compute the probability of no accidents per month
   c. Compute the probability of exactly one accident per month
   d. Compute the probability of fewer than three accidents per month.
   e. Computer the probability of three or more accidents per month.
5. In San Francisco, 30% of workers take public transportation daily.
   a. In a sample of 10 workers, what is the probability that exactly three workers take public transportation daily?
   b. What is the probability that less than two workers take public transportation daily?
6. A flight from Chicago to New York can take from 120 to 140 minutes. Assuming a uniform distribution, what is the probability that the flight will take less than 125 minutes?
7. A fair coin is tossed 10 times.
   a. What is the probability that you get exactly 4 heads?
   b. What is the probability that you get no more than 4 heads?
   c. What is the expected number of heads?
   d. What is the standard deviation?
8. The game of Yahtzee consists of five dice. A “Yahtzee” is getting the same number on all five dice.
   a. What is the probability that you get five sixes?
   b. What is the probability that you get a Yahtzee?
9. On average, 5 customers enter a specialty store every hour.
   a. What distribution would you use to model this?
   b. What is the probability that exactly 3 customers enter the store in the next hour?
   c. What is the probability that at least 3 customers enter the store in the next hour?
10. The average light bulb has a lifetime of 1000 hours.
    a. What distribution would you use to model lifetimes of light bulbs?
    b. What is the probability that a lightbulb fails after 500 hours?
    c. What is the probability that a lightbulb lasts between 1000 and 1500 hours?
    d. The manufacturer wants to provide a money-back guarantee if a lightbulb fails prematurely. If the manufacturer is willing to replace no more than 5% of all lightbulbs, he should replace any lightbulb which fails before how many hours of use?
11. At a certain university, 74% of female students own cars, while only 64% of male students do. 25% of the student body is female. A parking enforcement officer found a student vehicle parked illegally. What is the probability that the car was owned by a female student?

Chapter 3 - Inferential Statistics

When we study probability, we assume the general and draw conclusions about the particular. Thus, we make assumptions about the population (general), and estimate the probability of a sample (particular). The field of statistics is the inverse of this. Statistics involves collecting data (the particular) and drawing inferences about the population (general). The following chart illustrates this:

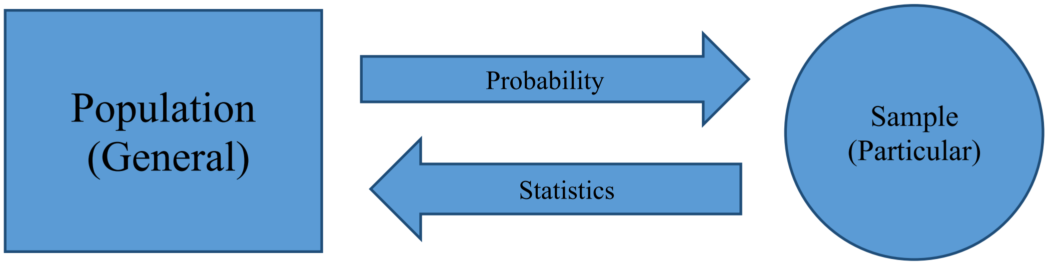


Figure 3 - Relationship between Probability and Statistics

3.1 Sampling Distributions

Let's take a random sample from any population, and estimate its mean. Then let us repeat this experiment over and over again. The resulting distribution has three properties:

1. The resulting distribution is normal for a sample size greater than 30.
2. The mean is the same as the population mean
3. The standard deviation equals the standard deviation divided by the square root of the sample size

Property number 1 is not intuitive. This means that regardless of the shape of the original distribution, the sampling distribution will always be bell-shaped for a reasonably large sample size.

Property number 2 suggests that the sample is unbiased. Let us take n random variables from any distribution and calculate the sample mean:

$$E(\bar{X}) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}(n\mu) = \mu$$

Property number 3 requires some explanation. The larger the sample size, the closer the sample mean is to the population mean. What this simply means is that more experienced people have a better picture of reality than do less experienced people. Would you rather have a surgery from a doctor who had performed the operation five times or one who had performed it 100 times? I think the answer is obvious.

$$Var(\bar{X}) = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}Var\sum_{i=1}^{n} X_i = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

The standard deviation of the sampling distribution is the square root of the variance and is known as the **standard error**.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Let's look at some examples:

Let us take 100 random samples of 30 heights from a hypothetical population of students with μ = 68 and $\sigma = 3$ and calculate their means:

```
      X←{mean 68 3 normal randomVariable ω}¨100/30
      mean X
68.026
```

The true standard error is:

```
      3 ÷ sqrt 30
0.54772
```

Notice that the following expression is very close to the true standard error.

```
      sdev X
0.52594
```

Notice that approximately 95 out of 100 samples fall within 1.96 standard deviations:

```
    sum 1.96>|(X-68.026) ÷ 0.52594
95
```

3.2 Confidence Intervals

When we estimate parameters from a sample, the result we get is usually close but incorrect. We would like to measure how close our estimate is; in fact, we are able to find an interval which usually contains the parameter. From the Central Limit Theorem, we know that the sampling distribution of the mean follows a normal distribution.

3.2.1 Means

The sample mean from our student data is shown below:

```
      mean #.SD.Height
68.776
```

But this value is probably not the same as the population height of all the students. But we can find an interval which most likely contains the true population mean. The sampling distribution for the mean tends to follow a normal distribution, thus $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. Solving for the mean gives: $\mu = \bar{X} - Z\frac{\sigma}{\sqrt{n}}$. Since the standard normal random variable is symmetric, we can find our confidence interval:

$$\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

We find the confidence interval for the height $CI = \bar{x} \mp Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$:

```
     XBAR(-,+)Z × SX ÷ sqrt N
68.406 69.147
```

Unfortunately, this formula assumes we know the population standard deviation σ. We could not possibly know this if we don't know the population mean. So we substitute the sample mean and perform the calculation.

Do you think Statistics is a dry subject? In the 1890's William Gossett worked for the Guinness Brewing Company. He dealt with small samples typically $n = 4$. (After all if one tried large samples of beer he wouldn't be in any condition to test for quality!) The problem was that the estimates for variance for samples of this size are relatively inaccurate. As an example, let us generate 100 samples of size 4 and calculate both their means and standard deviations:

```
   XXX←normal randomVariable each 100/4   ⍝ 100 samples of size 4
   M←mean each XXX                        ⍝ 100 sample means
   S←sdev each XXX                        ⍝ 100 sample standard deviations
   Z←(M-0) div (S ÷ 2)                    ⍝ This distribution should be N(0,1)
   normal criticalValue < 0.05 ÷ 2
1.96
   sum Z between ¯1.96 1.96           ⍝ We should get 95 but we only get 83!!
83
```

Clearly something is not right. Gossett realized that and found that if sigma is not known, the sampling distribution has fatter tails than the normal distribution. Since Guinness did not want him to publish trade secrets, Gossett published his results under the pseudonym “Student.” This distribution with fatter tails is called the Student t-Distribution or simply the t distribution. The distribution has one parameter, degrees of freedom. For a sample of size n, we need the t-distribution with $n - 1$ degrees of freedom. For our samples of size 4, we use $df = 3$. So if we want to contain 95% of all samples we calculate the critical value of the t-distribution:

```
   3 tDist criticalValue = .95
3.1824
```

Now using the critical value of the t-distribution, we do in fact capture 95 % of all samples:

```
   sum Z between ¯3.1824,3.1824
95
```

Thus for small samples with σ unknown, we use the following confidence interval:

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

The steps involved in creating a confidence interval (for σ unknown) are as follows:

Step	**TamStat Expression**

Estimate the sample mean: $\bar{x} = \frac{\sum x}{n}$	`XBAR←mean #.SD.Height`
	`68.776`

Estimate the standard error: $s_{\bar{X}} = \sqrt{\frac{s}{n}}$	`SX ← (sdev #.SD.Height)div sqrt count #.SD.Height`
	`0.72614`

Find the critical Value: $t_{\alpha/2}$	`T ← (N-1) tDist criticalValue < 0.05 div 2`
	`2.0262`

Find the Margin of Error: $E = t_{\alpha/2}s_p$	`E ← SX times T`
	`1.4713`

Calculate the confidence Interval: $\bar{x} \mp E$	`XBAR(-,+)E`
	`67.305 70.247`

The `confInt` operator combines all the steps above into one function. Notice that interval is wider when we use the t-distribution:

```
     mean confInt #.SD.Height
67.305 70.248
```

We are 95% confident that the mean height of the population is between 67.305 and 70.248 inches. What if we wanted to be more confident, like 99%? We then create a 99% confidence interval:

```
     .99 mean confInt #.SD.Height
66.805 70.748
```

Notice that this interval is wider. We are more confident but less accurate. We could reduce our confidence level to 90% to bring the interval even closer in:

```
     .90 mean confInt #.SD.Height
67.551 70.001
```

We can look at all three intervals simultaneously by listing all three confidence levels:

```
     .9 .95 .99 mean confInt #.SD.Height
67.5512546  70.00137698
67.30502236 70.24760922
66.80455688 70.7480747
```

To show the intervals graphically, use the `show` operator:

```
          .9 .95 .99 mean confInt show #.SD.Height
90%  (67.551,70.001)                               (------------------------*-----------------------)
95%  (67.305,70.248)                          (-----------------------------*----------------------------)
99%  (66.805,70.748)                (---------------------------------------*--------------------------------------)
                    +---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
                           66.5      67.0      67.5      68.0      68.5      69.0      69.5      70.0      70.5      71.0
```

So there is a trade-off. The more confident we are of our estimate, the less accurate and wider our interval. How can we both increase our confidence and narrow the interval? We must increase the sample size. Suppose we would like the margin of error to be within one inch of the true height at 95% confidence?

```
     S←sdev #.SD.Height
4.4762
     .95 mean sampleSize 1 S
80
```

Our original sample size was:

```
    count #.SD.Height
38
```

So we need to increase our sample size from 38 to 80. If we want 99% confidence we would have to increase our sample size to 135:

```
    .99 mean sampleSize 1 S
137
```

3.2.2 Proportions

A poll of likely voters is a sample of an election. We can estimate the percentage of voters who will vote for a particular candidate, or the percentage of voters who belong to a particular party. Let's take the student data and estimate the percentage of students who are Republicans. The point estimate is based on the formula:

$$\hat{p} = \frac{x}{n} = \frac{\text{Number of Republicans in Sample}}{\text{Sample Size}}$$

```
      P ← proportion #.SD.Party eq 'R'
      P
0.42105
```

We can construct a confidence interval around this estimate by using the fact that we can approximate the binomial distribution with the normal distribution. We can represent the random variable X (number of successes) in a sample of *n* likely voters by the binomial distribution with parameters *n* and π. The estimated proportion *p* is a random variable $= \frac{X}{n}$. From this we can calculate the expected value and variance of the sampling distribution *p*:

$$E\left[\frac{X}{n}\right] = \frac{1}{n}E[X] = \frac{1}{n}(np) = p \qquad \text{and} \qquad \text{Var}\left[\frac{X}{n}\right] = \frac{1}{n^2}\text{Var}[X] = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

We approximate the distribution of $\hat{p}$ with the normal distribution with $E(\hat{p}) = p$ and $Var(\hat{p}) = \frac{p(1-p)}{n}$ The steps involved:

Step	TamStat Expression

Estimate the sample proportion:
$p = \frac{x}{n}$
	`P ← proportion #.SD.Party eq 'R'`
	`0.42105`

Estimate the standard error:
$s_p = \sqrt{\frac{p(1-p)}{n}}$
	`sqrt  (P×1-P) div count #.SD.Party`
	`0.080093`

Find the critical Value:
$Z_{\alpha/2}$
	`Z ← normal criticalValue < 0.05 div 2`
	`1.96`

Find the Margin of Error:
$E = Z_{\alpha/2}s_p$
	`E ← SP times Z`
	`0.156979`

Calculate the confidence Interval: $p \mp E$	`P(-,+)E`
	`0.26407 0.57803`

In TamStat we can use the `confInt` operator:

```
    proportion confInt #.SD.Party eq 'R'
0.26407 0.57803
```

Obviously this interval (26.4% to 57.8%) is much too wide to be of any practical use. Again we can improve the accuracy by reducing the confidence level to 90%. We can show how this narrows the interval:

```
       .9 .95  proportion confInt show #.SD.Party eq 'R'
90%  (0.28931,0.55279)                  (-------------------------*--------------------------)
95%  (0.26407,0.57803)             (------------------------------*-------------------------------)
                      +---------+---------+---------+---------+---------+---------+---------+---------+---------+---------
                             0.25      0.30      0.35      0.40      0.45      0.50      0.55      0.60      0.65      0.7
```

This interval is still too wide. We must increase the sample size, but how large should it be? Suppose we would like to be with 3% of the true proportion. We will start with the formula for the margin of error and solve for the sample size *n*. Thus

$$E = Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \qquad \text{yields} \qquad n = p(1-p)\left(\frac{Z_{\alpha/2}}{E}\right)^2.$$

Applying the formula:

```
      P × (1-P) × (Z ÷ 0.03)*2
1040.5
```

We can also use the `sampleSize` operator which rounds up the sample to the next higher integer:

```
      .95 proportion sampleSize .03 P
1041
```

The sample size for a 99% confidence interval would be:

```
        .99 proportion sampleSize .03 P
1798
```

One of the problems with this formula is that it requires knowledge about the proportion which is what we are trying to find in the first place. If we have no prior knowledge of the proportion, to be safe we should find the value of *p* which yields the largest sample size.

Let $n(p) = p(1-p)\left(\frac{Z_{\alpha/2}}{E}\right)^2 = \left(\frac{Z_{\alpha/2}}{E}\right)^2 (p - p^2)$

To find the maximum we set $n'(p) = \left(\frac{Z_{\alpha/2}}{E}\right)^2 (1-2p) = 0$. This requires that $p = \frac{1}{2}$. We can show this a maximum by showing $n"(p) = -2\left(\frac{Z_{\alpha/2}}{E}\right)^2 < 0$ since the squared value is positive.

Thus if we have no clue as to what *p* should be, use the value 0.5:

```
     .95 proportion sampleSize .03 0.5
1068
```

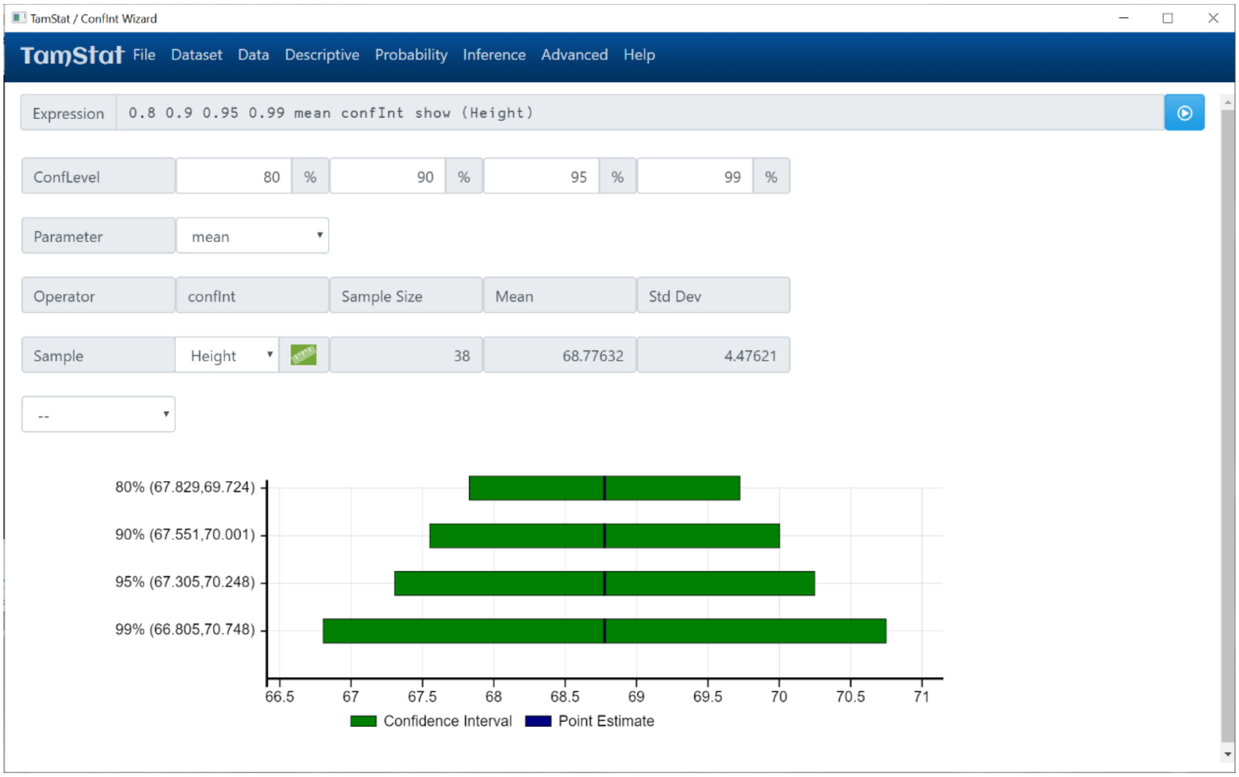
That is why most polls use a random sample of about 1000 voters so that the margin of error is +/- 3 percentage points.

3.2.3 Confidence Intervals Using the Wizard

From the main menu select "Inference", then select "Confidence Intervals". Select the appropriate parameter, e.g. mean or proportion. You may select an appropriate variable or select "stats" if you know the sample size, mean and standard deviation to calculate the confidence interval for the mean, or the sample size and number of successes (events). For proportions, you may also use an expression resulting in a Boolean vector (all 1's and 0's). You may select one or more confidence levels as indicated on the top row.

To calculate the confidence interval, select "calculate test statistic". The correct TamStat expression will appear in the expression box, and the report will contain the appropriate confidence intervals alongside a graphic showing the various intervals along a number line.

The screens appear on the following page:



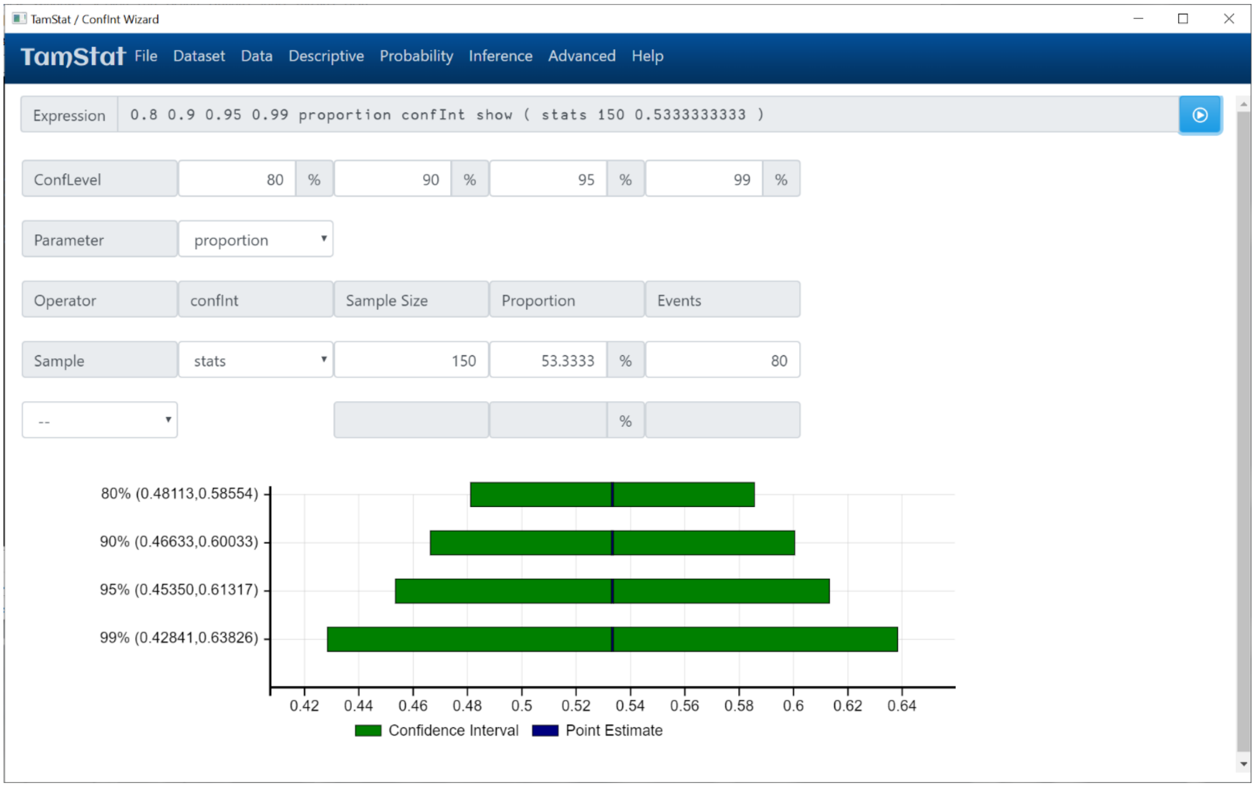
Expression `0.8 0.9 0.95 0.99 mean confInt show (Height)`

ConfLevel 80 % 90 % 95 % 99 %

Parameter mean

Operator confInt Sample Size Mean Std Dev

Sample Height 38 68.77632 4.47621



Expression `0.8 0.9 0.95 0.99 proportion confInt show ( stats 150 0.5333333333 )`

ConfLevel 80 % 90 % 95 % 99 %

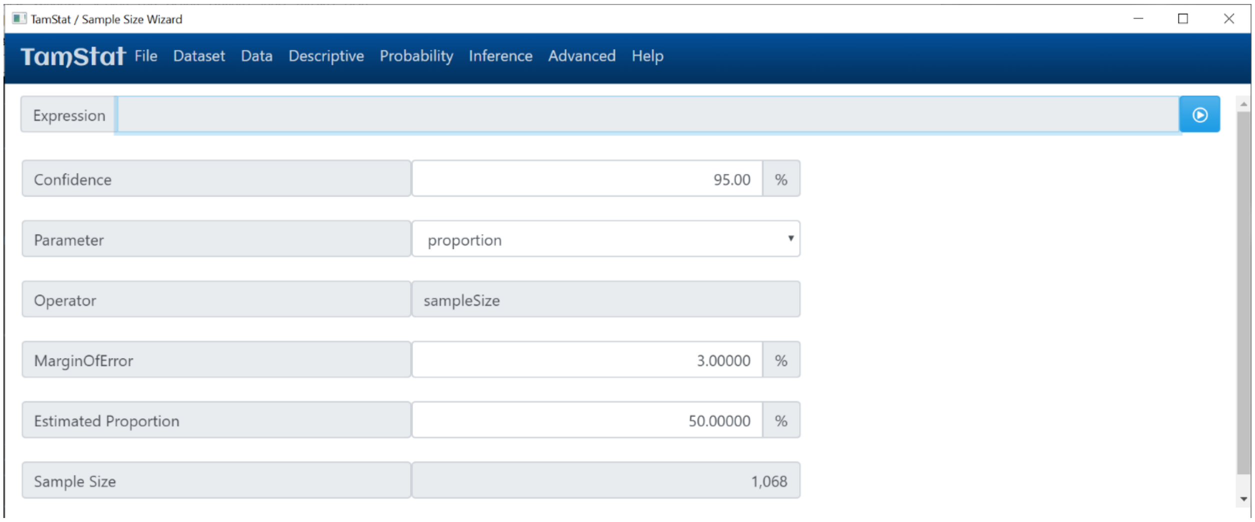
Parameter proportion

Operator confInt Sample Size Proportion Events

Sample stats 150 53.3333 % 80

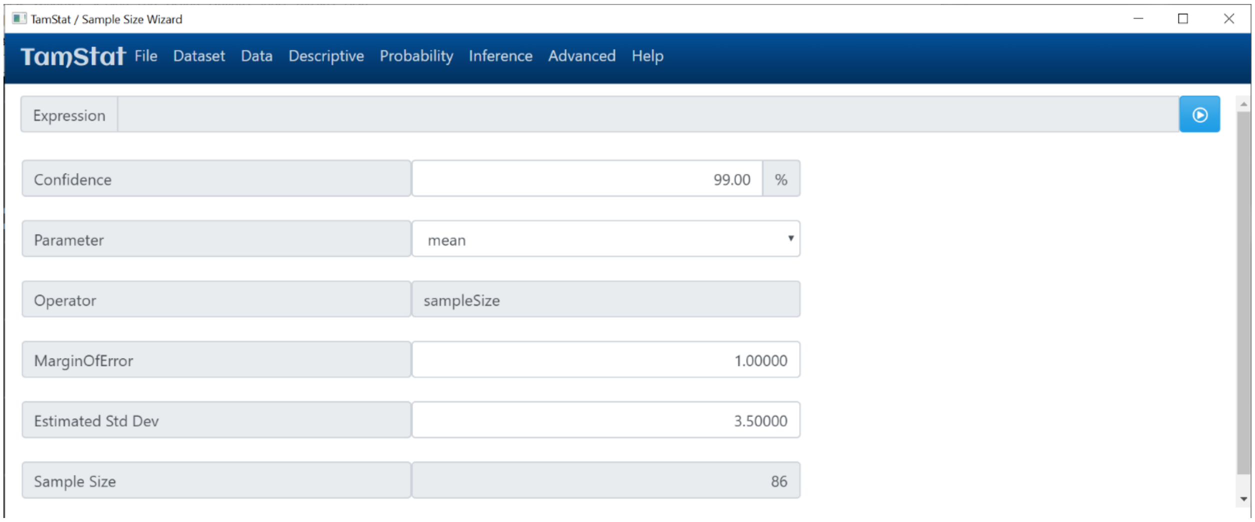
3.2.4 Sample Size Wizard

To determine the correct sample size to use for a test or confidence interval, select "Inference", then select "Sample Size". The following screen will appear:


TamStat / Sample Size Wizard
TamStat File Dataset Data Descriptive Probability Inference Advanced Help

Expression		
Confidence	95.00	%
Parameter	proportion ▾	
Operator	sampleSize	
MarginOfError	3.00000	%
Estimated Proportion	50.00000	%
Sample Size	1,068	

Enter the appropriate confidence level, parameter, margin of error, with an optional estimate for the proportion and the standard deviation for the mean. The sample size will appear next to the words "Sample Size".


TamStat / Sample Size Wizard
TamStat File Dataset Data Descriptive Probability Inference Advanced Help

Expression		
Confidence	99.00	%
Parameter	mean ▾	
Operator	sampleSize	
MarginOfError	1.00000	
Estimated Std Dev	3.50000	
Sample Size	86	

3.3 Hypothesis Tests

In a court of law, a person is considered not guilty unless there is reasonable doubt. Hypothesis testing is similar in that we make an assumption as to the value of a parameter. Since we cannot determine the exact value, we assume it is true unless there is significant evidence to the contrary. In a trial, there can be four possible outcomes:

	Verdict	
The Truth	Guilty	Not Guilty
Defendant Is Guilty	OK $(1-\beta)$	Type II Error β
Defendant innocent	Type I Error α	OK $(1-\alpha)$

In the above chart, we hope that the probability of a type I is very small. Unfortunately, this probability can be zero. A typical value is $\alpha = 0.05$ but it could also be 0.1 or 0.01 for example. We would also like β to be small, but not at the risk of increasing α.

In the above chart, we assume that the defendant is innocent. We cannot prove this; the burden of proof is on the prosecution who must gather enough evidence to prove guilt; this is known as the alternative hypothesis.

The null hypothesis assumes the relationship of a parameter to a particular value. The alternative hypothesis is the complement of the null hypothesis. Both cannot be true, but one must be true.

3.3.1 Means

Let us assume that the average weight of students is 150 pounds. The null hypothesis represents our assumption:

$H_0: \mu = 150$

The alternative hypothesis is the complement:

$H_1: \mu \neq 150$

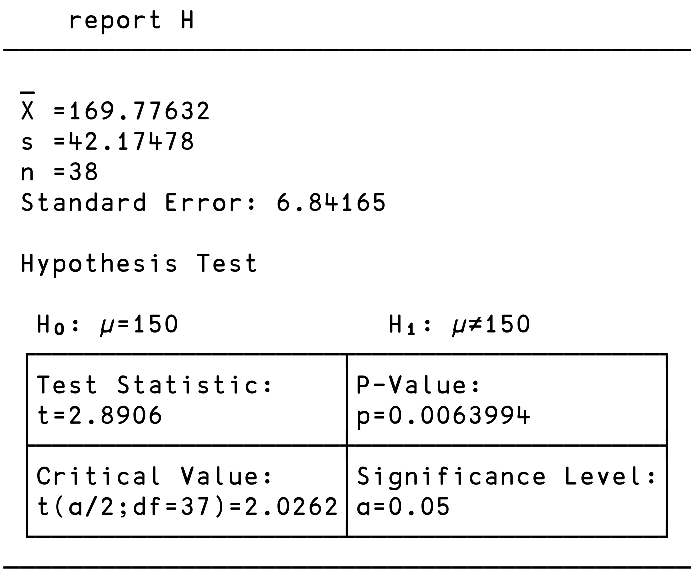
We can run the hypothesis operator assigning mean to the left operand and = to the right operand:

```
H←#.SD.Weight mean hypothesis = 150
```

The result is a namespace, which contains various results. Two of the most important are the test statistic and the p-value:

```
      H.TestStatistic
2.8906
      H.P
0.0063994
```

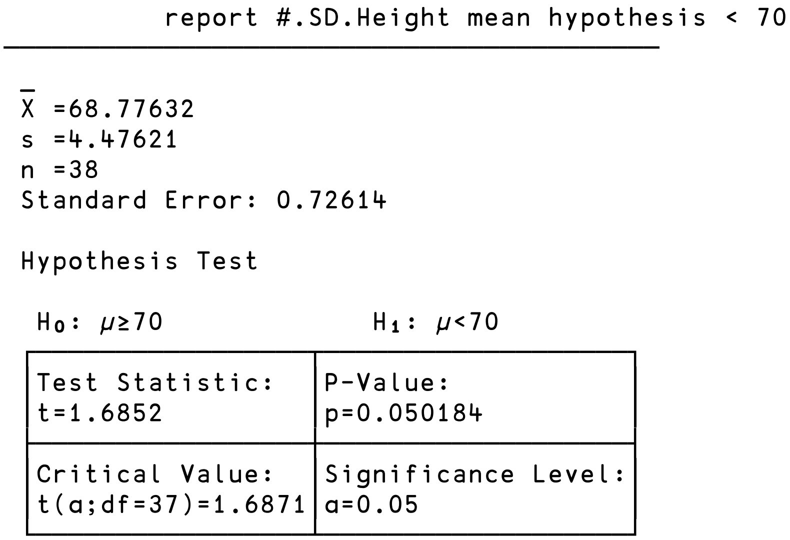
However, it is much easier to simply generate a report from the namespace:



```
    report H
___________________________________________

 _
 X =169.77632
 s =42.17478
 n =38
 Standard Error: 6.84165

 Hypothesis Test

  H0: μ=150              H1: μ≠150
 ┌────────────────────┬────────────────────┐
 │Test Statistic:     │P-Value:            │
 │t=2.8906            │p=0.0063994         │
 ├────────────────────┼────────────────────┤
 │Critical Value:     │Significance Level: │
 │t(α/2;df=37)=2.0262 │α=0.05              │
 └────────────────────┴────────────────────┘
___________________________________________
```

Since the p-value is less than the significance level, and the test statistic is greater than the critical value, we reject the null hypothesis. Thus the sample mean is shown to be significantly different than 150 lbs.

Now we would like to test whether the average student height is significantly less than 70 inches. Our claim is > 70 . The null hypothesis must contain equality. Therefore we must assign our claim to the alternate hypothesis; the counter claim $\mu \geq 70$ becomes the null hypothesis.



```
          report #.SD.Height mean hypothesis < 70
_________________________________________

 _
 X =68.77632
 s =4.47621
 n =38
 Standard Error: 0.72614

 Hypothesis Test

  H0: μ≥70              H1: μ<70
 ┌──────────────────┬───────────────────┐
 │Test Statistic:   │P-Value:           │
 │t=1.6852          │p=0.050184         │
 ├──────────────────┼───────────────────┤
 │Critical Value:   │Significance Level:│
 │t(α;df=37)=1.6871 │α=0.05             │
 └──────────────────┴───────────────────┘
```

In this case the p-value is slightly greater than the significance level and test statistic is slightly lower than the critical value, so we fail to reject the null hypothesis.

3.3.2 Proportions

We can test a hypothesis test for proportions. Let's test whether the proportion of students who are Republicans is less than 50%.

```
     report (#.SD.Party eq 'R')proportion hypothesis < 0.5
```

```
 p =0.42105
 n =38
 Standard Error: 0.08111

 Hypothesis Test

  H₀: p≥0.5            H₁: p<0.5
 ┌────────────────┬───────────────────┐
 │Test Statistic: │P-Value:           │
 │Z=0.97333       │p=0.16519          │
 ├────────────────┼───────────────────┤
 │Critical Value: │Significance Level:│
 │Z(α)=1.6449     │α=0.05             │
 └────────────────┴───────────────────┘
```

We don't have enough evidence to reject the null hypothesis, so we can't conclude that the proportion of students who are Republicans is less than 50%.

3.4 Two Sample Tests

3.4.1 Difference of Means (Independent Samples)

To find the difference between male height and female height, we can generate two separate groups of data:

```
      MALE ←  68 72 69 71 65
      FEMALE ←  62 66 65 64
      mean confInt MALE FEMALE
1.189210037 8.310789963
```

Another way to do this is to find a numeric variable and use a Boolean variable of the same length to separate it into two groups.   We find the confidence interval for the difference showing male students average between 3.8 and 10.3 taller than female students:

```
     mean confInt #.SD.Height (#.SD.Sex eq 'M')
3.825085589 10.29368836
```

To obtain confidence intervals for several independent groups within the same variable, use a qualitative variable e.g.

```
      mean confInt #.SD.Height #.SD.Sex
69.27328612 71.6232656
60.26785892 66.50991886
```

To identify the two groups, use the `show` operator:

```
      mean confInt show #.SD.Height #.SD.Sex
```

```
95% M(69.273,71.623)                                             (-----*-----)
95% F(60.268,66.510) (---------------*---------------)
                    +---------+---------+---------+---------+---------+---------+---------+
                             62        64        66        68        70        72        74
```

Now let us perform a hypothesis test on the same data:

```
  (MaleHeight FemaleHeight) ← #.SD.Height splitBy #.SD.Sex eq 'M'

    report MaleHeight mean hypothesis = FemaleHeight
```

```
 _                    _
 X1=70.44828          X2=63.38889
 s1=3.08899           s2=4.06031
 n1=29                n2=9
 Standard Error: 1.46997

 Hypothesis Test

  H0: μ1=μ2              H1: μ1≠μ2
 ┌───────────────────┬────────────────────┐
 │Test Statistic:    │P-Value:            │
 │t=4.8024           │p=0.00055123        │
 ├───────────────────┼────────────────────┤
 │Critical Value:    │Significance Level: │
 │t(α/2;df=11)=2.201 │α=0.05              │
 └───────────────────┴────────────────────┘
```

We reject the hypothesis that male and female heights are equal. Notice that zero is not in the confidence interval that we calculated previously, since both bounds are positive. This is consistent with the results of the hypothesis. Two means not being equal is equivalent to their difference not being zero.

3.3.5 Mean of the Differences (Paired Data)

Some courses offer a pre-test and a post-test to show the effectiveness of the course. The same six people take the test before and after completing the course. The results are shown here:

```
      BEFORE  ←  54 49 68 66 62 62
      AFTER   ←  50 65 74 64 68 72
```

Here, the data are paired; the first score in the BEFORE vector corresponds to the first score in the AFTER vector. Thus we can subtract one from the other without getting a length error and simply calculate a confidence interval on the mean of the differences:

```
      mean confInt AFTER-BEFORE
¯2.4824 13.149
```

Here the confidence interval straddles zero since the lower bound is negative and the upper bound is positive.

We can also perform a hypothesis test on the differences. We apply the paired operator to the mean to accomplish this:

```
     report AFTER mean paired hypothesis > BEFORE
```

```
 _
 d =5.33333
 sd =7.44759
 n =6
 Standard Error: 3.04047

 Hypothesis Test
```

```
  H₀: μd≤0                 H₁: μd>0 (Claim)
┌──────────────────────┬──────────────────────┐
│Test Statistic:       │P-Value:              │
│t=1.754116039         │p=0.06989             │
├──────────────────────┼──────────────────────┤
│Critical Value:       │Significance Level:   │
│t(α;df=5)=2.015048358 │α=0.05                │
└──────────────────────┴──────────────────────┘
  Conclusion: Fail to reject H₀
──────────────────────────────────────────────
```

Although the average improvement is over 5 points, it is not significant; therefore we cannot conclude that the test is effective.

3.5 Inferences about Population Variances

3.5.1 Confidence Intervals

To calculate a confidence interval for the variance, we use the Chi-Square distribution. If the sample variance from a population with a normal distribution with true variance of σ^2 is s^2 , the following has a chi-square distribution:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Solving for σ^2 gives us the lower and upper confidence bounds:

$$\frac{(n-1)s^2}{\chi^2_L} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_U}$$

The steps involved in creating a confidence interval (for σ unknown) are as follows:

Step	TamStat Expression

Estimate the sample variance:
$s^2 = \frac{\sum(x-\bar{x})^2}{n}$
	`      var #.SD.Height`
	`20.036`

Find the degrees of freedom
$n-1$
	`      (count #.SD.Height)-1`
	`37`

Find the critical Values:
$\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$
	`      X2←37 chiSquare criticalValue < .025 .975`
	`55.668 22.106`

Calculate the confidence Interval:	`      37 times 20.036 div X2`
	`13.317 33.536`

For standard deviation, take the square root	`      sqrt 13.317 33.536`
	`3.6492 5.791`

Of course we can simply run the confidence interval on the variance or the standard deviation:

```
      var confInt #.SD.Height
13.317 33.537
      sdev confInt #.SD.Height
3.6493 5.7911
```

3.5.2 Hypothesis Tests

Most hypothesis tests for variance are upper tail. This is because we are not usually concerned about the lower tail, since smaller variances are not usually a problem. Thus, to test that the variance of student heights is not more than 20, we do the following:

```
      report #.SD.Height var hypothesis > 20
```

```
 s² =20.03645
 n =38

 Hypothesis Test

  H₀: σ²≤20             H₁: σ²>20
 ┌───────────────────┬────────────────────┐
 │Test Statistic:    │P-Value:            │
 │χ²=37.067          │p=0.46565           │
 ├───────────────────┼────────────────────┤
 │Critical Value:    │Significance Level: │
 │χ²(α;df=37)=52.192 │α=0.05              │
 └───────────────────┴────────────────────┘
```

In the above example, we fail to reject the null hypothesis because the p-value is greater than 0.05.

Since variances are in squared units, we may prefer to work with the standard deviation. This is not a problem; we simply replace the operand "`var`" with "`sdev`". We also need to specify units, not squared units:

```
   report #.SD.Height sdev hypothesis > 4
```

```
 s =4.47621
 n =38

 Hypothesis Test

  H₀: σ≤4               H₁: σ>4
 ┌───────────────────┬────────────────────┐
 │Test Statistic:    │P-Value:            │
 │χ²=46.334          │p=0.13978           │
 ├───────────────────┼────────────────────┤
 │Critical Value:    │Significance Level: │
 │χ²(α;df=37)=52.192 │α=0.05              │
 └───────────────────┴────────────────────┘
```

First create two populations. Let us compare whether the variances of heights between male and female students are different.

```
      MaleHeight FemaleHeight ← #.SD.Height splitBy #.SD.Sex eq 'M'
```

Then run the hypothesis test comparing Male Height to Female Height:

```
     report MaleHeight var hypothesis eq FemaleHeight
```

```
s^2_1=9.54187           s^2_2=16.48611
n_1=29                 n_2=9

Hypothesis Test

 H_0: σ_1^2=σ_2^2             H_1: σ_1^2≠σ_2^2
+----------------------+--------------------+
|Test Statistic:       |P-Value:            |
|F=0.57878             |p=0.86436           |
+----------------------+--------------------+
|Critical Value:       |Significance Level: |
|F(α/2;df=28 8)=3.9093 |α=0.05              |
+----------------------+--------------------+
```

Notice that the hypothesis operator uses the F-Distribution instead of the chi-Square distribution to compare two variances. The same rules apply here: Since the p-value is greater than the significance level of 0.05, we fail to reject the null hypothesis and conclude that the variances are equal.

3.5.3 Testing two means when variances are equal

In section 3.4.1 above, we made no assumption about the variances of each population. If the variances of two populations are equal, we can use more degrees of freedom for our two-sample t-test. The degrees of freedom when the variances are equal are: $df = n_1 + n_2 - 2$. We pool the variances using the following formula:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

As we saw in section 3.5.2, the variances for the male and female heights were not significantly different, so we can assume they are equal. In TamStat, we can do this by splitting the Height variable into two groups using the splitBy function.

```
        report mean hypothesis = #.SD.Height splitBy #.SD.Sex
```

```
X̄_1=70.44828          X̄_2=63.38889
s_1=3.08899           s_2=4.06031
n_1=29                n_2=9
Standard Error: 1.27040

Hypothesis Test

 H_0: μ_1=μ_2 (Claim)         H_1: μ_1≠μ_2
+--------------------------+--------------------+
|Test Statistic:           |P-Value:            |
|t=22.22730855             |p<0.00001           |
+--------------------------+--------------------+
|Critical Value:           |Significance Level: |
|t(α/2;df=36)=2.028093968  |α=0.05              |
+--------------------------+--------------------+
```

3.6 Generalized Hypothesis Test Procedures

In real life, it will not always be obvious which type of hypothesis test to apply. Are you dealing with means or proportions? Do you perform a one-sample or two-sample test? Do you want a lower-tail, upper-tail or two-tail test? To accomplish this, one must look for clues in the problem statement. Try the following 3-step procedure to conduct a hypothesis test. The appropriate formulas are listed on the following page.

3.6.1 Steps in Hypothesis Testing

All hypothesis tests can be processed using the following procedures (see next page):

Step 1. State the Null and Alternative Hypotheses:

a. Determine the problem type: mean, proportion, variance difference of means, difference of proportions, paired data.
b. Pick the relational symbol which best describes the claim:

= ≥ ≤ ≠ < >

Examples: changed ≠, decreased <, increased > stayed the same =, at least ≥
improved < or > depending on the context, e.g. for quality (proportion of defects) use <, for fuel efficiency (miles per gallon) use >.

c. Also pick its complement:

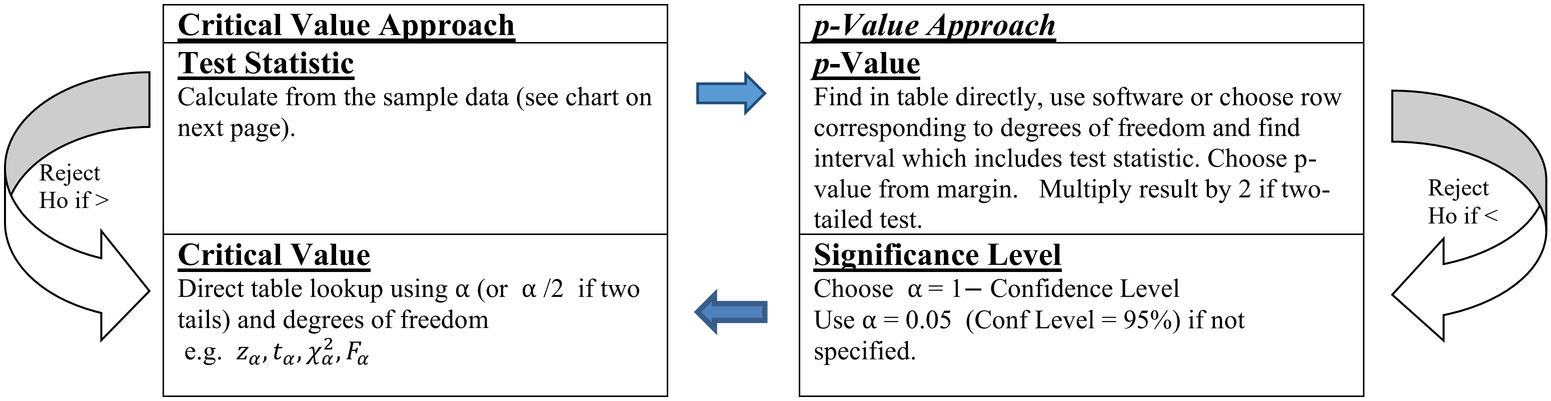
Symbol	=	≥	≤	≠	<	>
Complement	≠	<	>	=	≥	≤

d. From the pair of symbols that you selected, associate the relational symbol containing equality to the null hypothesis (Ho);
= ≥ ≤

e. Associate the remaining symbol to the alternate hypothesis (H_1).
≠ < >

Step 2: Complete the comparison table. Enter the significance level; then calculate the test statistic. Obtain the critical and p-values from the appropriate tables (or use software).

Comparison Table Layout



Critical Value Approach

Test Statistic
Calculate from the sample data (see chart on next page).

Critical Value
Direct table lookup using α (or α /2 if two tails) and degrees of freedom
e.g. $z_\alpha, t_\alpha, \chi^2_\alpha, F_\alpha$

***p*-Value Approach**

***p*-Value**
Find in table directly, use software or choose row corresponding to degrees of freedom and find interval which includes test statistic. Choose p-value from margin. Multiply result by 2 if two-tailed test.

Significance Level
Choose α = 1− Confidence Level
Use α = 0.05 (Conf Level = 95%) if not specified.

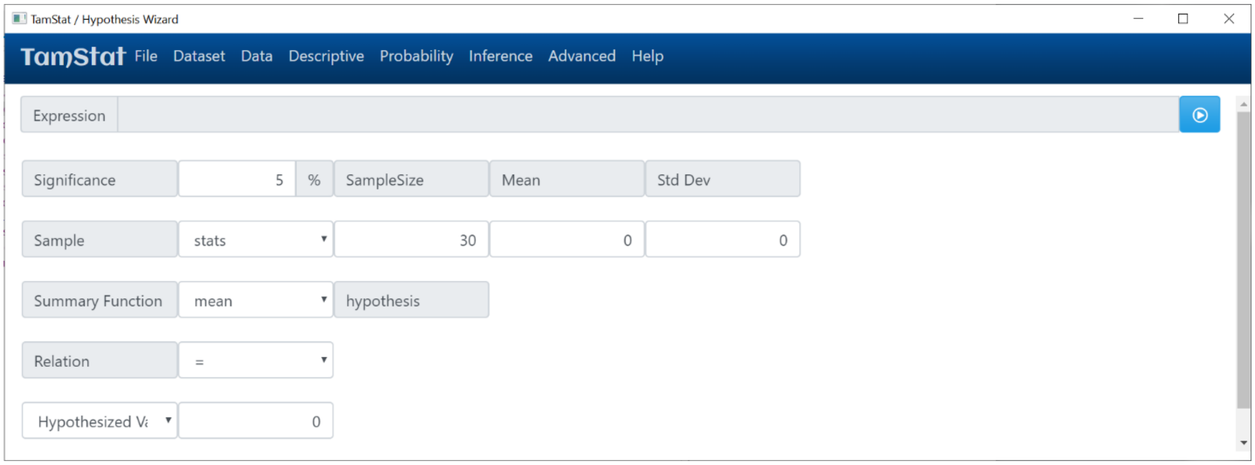
Step 3: Apply the appropriate rejection rules (p-value < α and/or test statistic > critical value) , either reject or do not reject Ho, and state your conclusion in words.

Test	H_1	(Adjusted) Test Statistic	Dist. and d.f.
Mean	$\mu \neq \mu_0$ $\mu < \mu_0$ $\mu > \mu_0$	$\lvert t \rvert$ $-t$ t $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	Student t , n-1
Proportion	$p = p_0$ $p < p_0$ $p > p_0$	$\lvert z \rvert$ $-z$ z $z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$; $\hat{p} = \frac{x}{n}$	Normal (∞)
Variance	$\sigma^2 \neq \sigma_0^2$ $\sigma^2 > \sigma_0^2$	$\frac{(n-1)s^2}{\sigma^2}$	Chi-Square (n-1)
Difference of Means	$\mu_1 \neq \mu_2$ $\mu_1 > \mu_2$	$\sigma_1 = \sigma_2$ $\lvert t \rvert$ t $t = \frac{\bar{x}_1 - \bar{x}_2}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$; $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$	Student t $n_1 + n_2 - 2$
	$\mu_1 \neq \mu_2$ $\mu_1 > \mu_2$	$\sigma_1 \neq \sigma_2$ $\lvert t \rvert$ t $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{v_1 + v_2}}$; $v_1 = \frac{s_1^2}{n_1}$ $v_2 = \frac{s_2^2}{n_2}$	Student t $\frac{(v_1 + v_2)^2}{\left(\frac{v_1^2}{n_1 - 1} + \frac{v_2^2}{n_2 - 1}\right)}$
Paired Difference	$\mu_d \neq 0$ $\mu_d > 0$	$\lvert t \rvert$ t $t = \frac{\bar{d}}{s_d/\sqrt{n}}$; $\bar{d} = \frac{\sum d}{n}$; $s_d = \sqrt{\frac{(d - \bar{d})^2}{n-1}}$ $d = x_1 - x_2$	$n-1$
Difference of Proportion	$p_1 = p_2$ $p_1 > p$	$\lvert z \rvert$ z $z = \frac{p_1 - p_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$; $p_1 = x_1/n_1$ $p_2 = x_2/n_2$ $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$	Normal Use ∞
Two Variances	$\sigma_1^2 \neq \sigma_2^2$ $\sigma_1^2 > \sigma_2^2$	$F = s_1^2 / s_2^2$	$F(n_1 - 1; n_2 - 1)$
Goodness of Fit	Not from distribution	$\chi^2 = \sum \frac{(O-E)^2}{E}$ $E = nP(x)$; for uniform $E = n/k$	$\chi^2 \quad (k-1)$
Independence	Not independent	$\chi^2 = \sum \frac{(O-E)^2}{E}$ $E = \frac{\text{ColumnTotal} \times \text{RowTotal}}{\text{GrandTotal}}$	$\chi^2 \ (r-1)(c-1)$
Correlation	$\rho \neq 0$	$\lvert \boldsymbol{t} \rvert$; $t = r \Big/ \sqrt{(1-r^2)/(n-2)}$	Student t (n-2)
Slope	$\beta_i \neq 0$	$\lvert \boldsymbol{t} \rvert$; $t = b_i / s_{b_i}$	Student t (n-k-1)

Conf Int	Sample Size	Standard Error	Margin of Error	Confidence Bounds
Mean	$n = \left\lceil (\sigma Z_{\alpha/2}/E)^2 \right\rceil$	$s_{\bar{x}} = s/\sqrt{n}$	$E = t_{\alpha/2} s_{\bar{x}}$	$\bar{x} \mp E$
Proportion	$n = \left\lceil p(1-p)(Z_{\alpha/2}/E)^2 \right\rceil$	$s_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}$	$E = z_{\alpha/2} s_{\hat{p}}$	$\hat{p} \mp E$

3.6.2 Examples of Hypothesis testing using the Wizard:

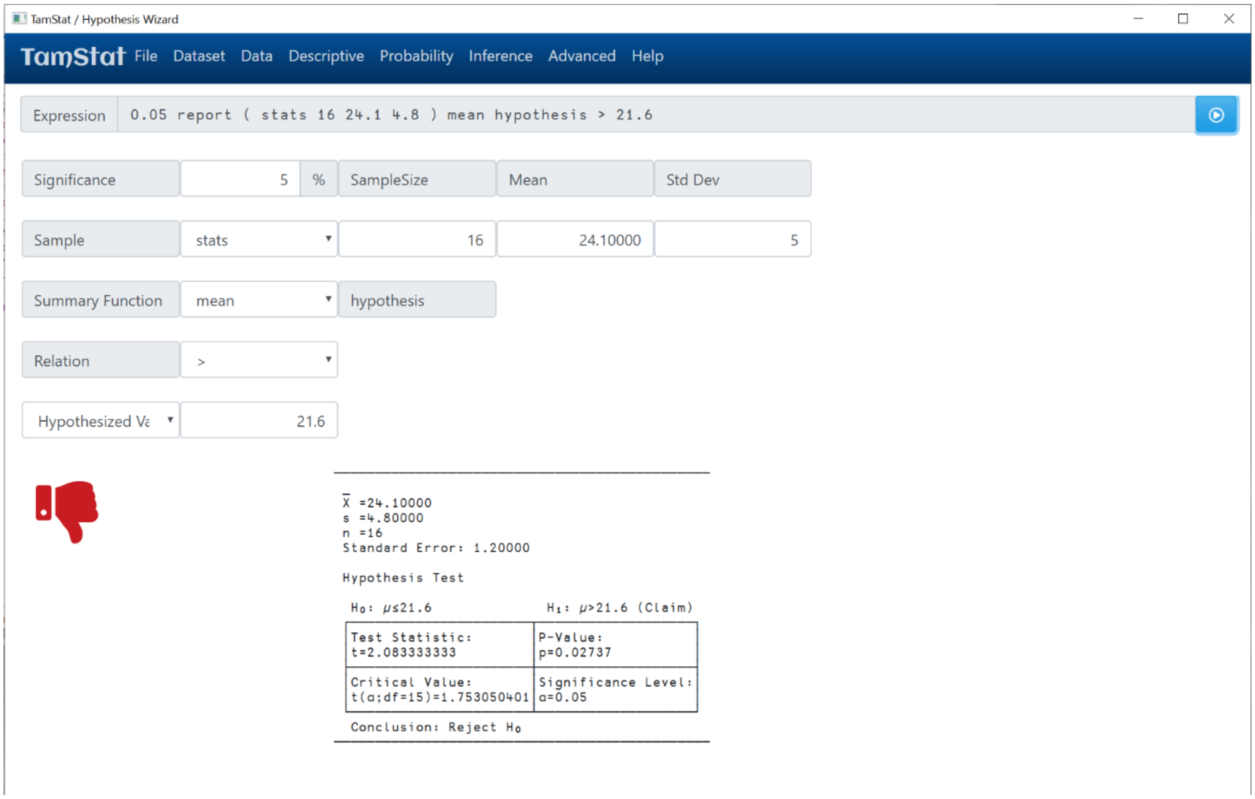
To use the Hypothesis Wizard, Select “Inference” from the main menu, then select “HypothesisTests” .
The following screen will appear:



Expression					
Significance	5	%	SampleSize	Mean	Std Dev
Sample	stats		30	0	0
Summary Function	mean		hypothesis		
Relation	=				
Hypothesized Va	0				

Annual per-capita consumption of milk is 21.6 gallons. A sample of 16 individuals from the the town of Webster showed a sample average consumption of 2.41 gallons with a standard deviation of 4.8 gallons. Test the hypothesis that Webster’s average milk consumption is higher than the national average.

Fill out the Wizard as indicated below and select “Calculate Test Statistic”:

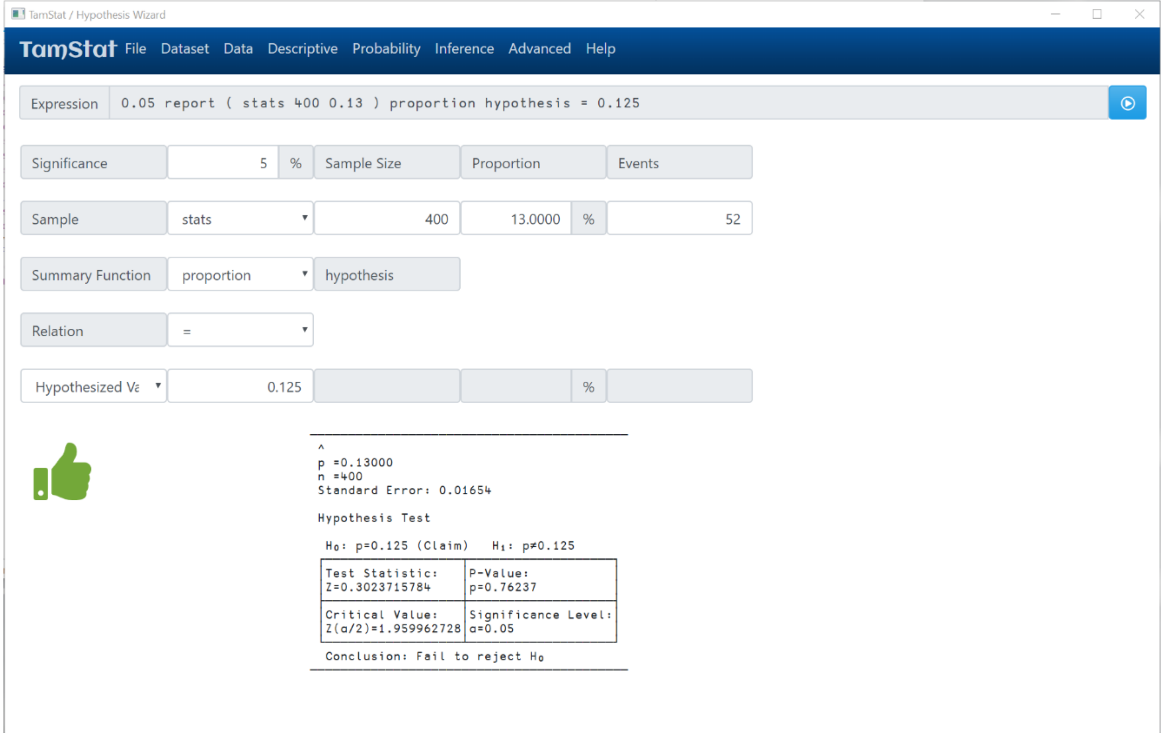


Expression	0.05 report (stats 16 24.1 4.8) mean hypothesis > 21.6				
Significance	5	%	SampleSize	Mean	Std Dev
Sample	stats		16	24.10000	5
Summary Function	mean		hypothesis		
Relation	>				
Hypothesized Va	21.6				

```
X̄ =24.10000
s =4.80000
n =16
Standard Error: 1.20000

Hypothesis Test

 H₀: μ≤21.6                  H₁: μ>21.6 (Claim)
┌────────────────────────┬────────────────────┐
│Test Statistic:         │P-Value:            │
│t=2.083333333           │p=0.02737           │
├────────────────────────┼────────────────────┤
│Critical Value:         │Significance Level: │
│t(α;df=15)=1.753050401  │α=0.05              │
└────────────────────────┴────────────────────┘
 Conclusion: Reject H₀
```

A recent study showed that 12.5% of American workers belonged to unions. In a sample of 400 American workers, 52 said they belonged to unions. Test whether union membership has increased. For “summary function”, select “proportion”.

Note that the last two columns are now labeled “Events” and “Proportion”. Enter “400” for sample size and “52” for events. The sample proportion will automatically be calculated.

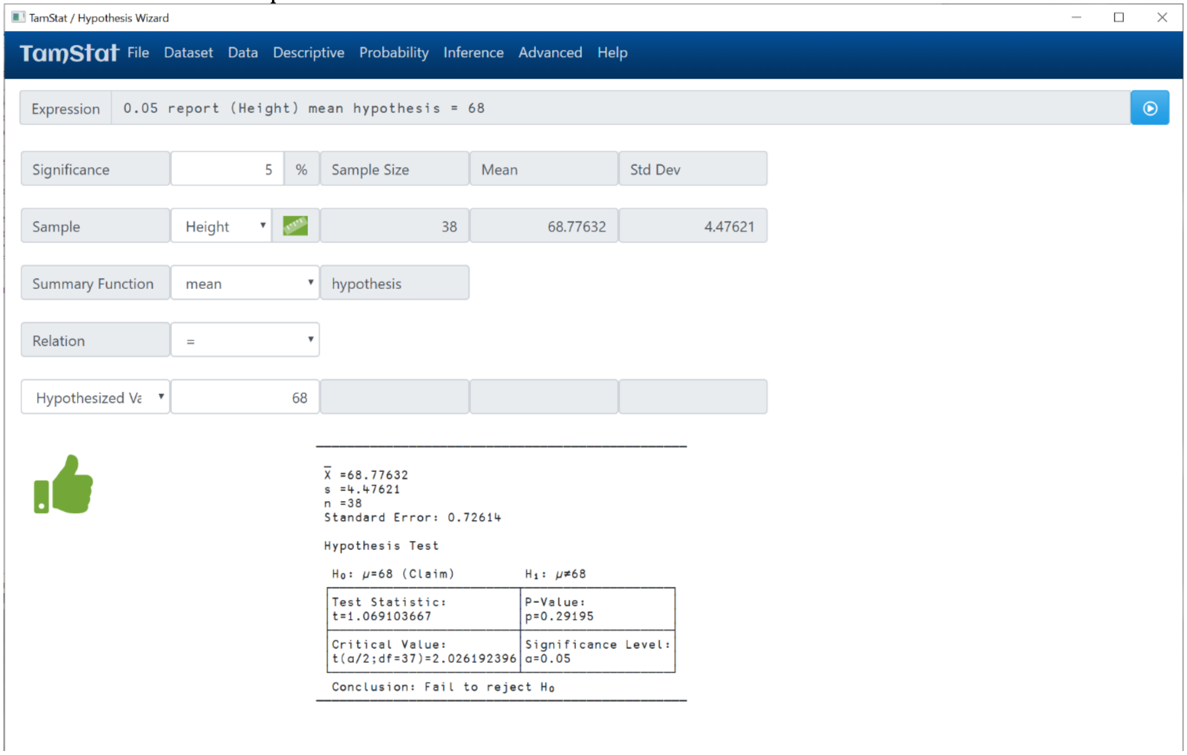


Expression `0.05 report ( stats 400 0.13 ) proportion hypothesis = 0.125`

Significance	5	%	Sample Size	Proportion		Events
Sample	stats		400	13.0000	%	52
Summary Function	proportion		hypothesis			
Relation	=					
Hypothesized Va	0.125				%	

```
^
p =0.13000
n =400
Standard Error: 0.01654

Hypothesis Test

 H0: p=0.125 (Claim)   H1: p≠0.125
┌──────────────────────┬──────────────────────┐
│Test Statistic:       │P-Value:              │
│Z=0.3023715784        │p=0.76237             │
├──────────────────────┼──────────────────────┤
│Critical Value:       │Significance Level:   │
│Z(α/2)=1.959962728    │α=0.05                │
└──────────────────────┴──────────────────────┘
 Conclusion: Fail to reject H0
```

To use data from a database, simply replace “stats” with the name of a variable. Selecting the drop-down combo will provide a list of all variables currently in the workspace including all database variables indicated by [database name].[variable name]. Test whether the average height of students is 68 inches. Note that TamStat calculates sample mean and standard deviation from the data.

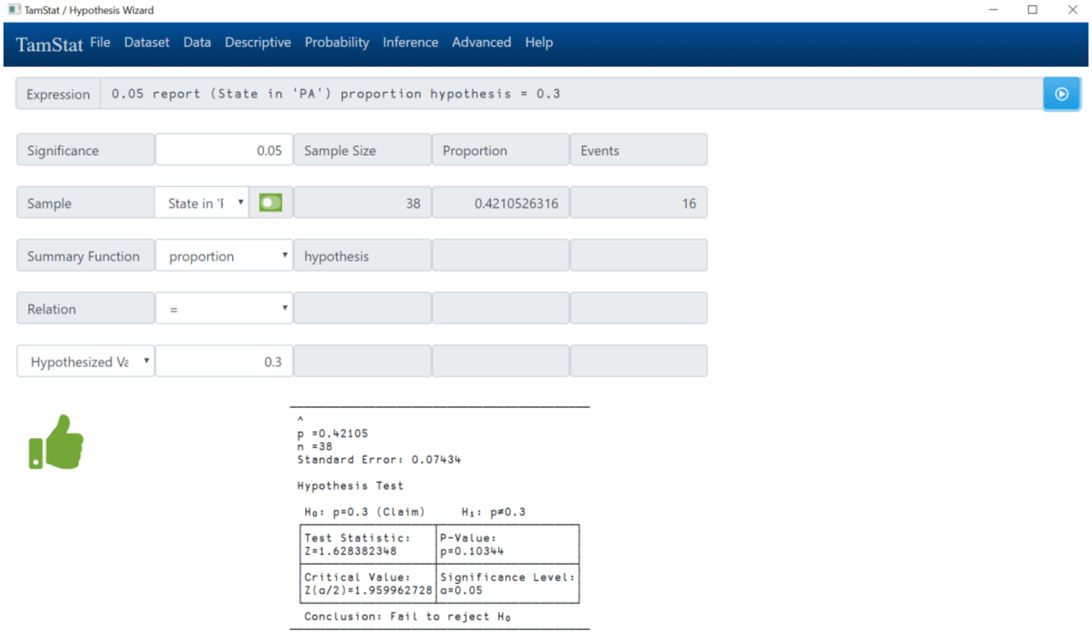


Expression `0.05 report (Height) mean hypothesis = 68`

Significance	5	%	Sample Size	Mean	Std Dev
Sample	Height		38	68.77632	4.47621
Summary Function	mean		hypothesis		
Relation	=				
Hypothesized Va	68				

```
X̄ =68.77632
s =4.47621
n =38
Standard Error: 0.72614

Hypothesis Test

 H0: μ=68 (Claim)          H1: μ≠68
┌──────────────────────────┬──────────────────────┐
│Test Statistic:           │P-Value:              │
│t=1.069103667             │p=0.29195             │
├──────────────────────────┼──────────────────────┤
│Critical Value:           │Significance Level:   │
│t(α/2;df=37)=2.026192396  │α=0.05                │
└──────────────────────────┴──────────────────────┘
 Conclusion: Fail to reject H0
```

At the 10% level of significance, test the hypothesis that 30% of students at the University of Scranton are from Pennsylvania: Here, instead of using a simple variable name `STATE`, we can use an expression, `STATE in 'PA'` which results in a Boolean vector, appropriate for testing proportions:


TamStat / Hypothesis Wizard
TamStat File Dataset Data Descriptive Probability Inference Advanced Help

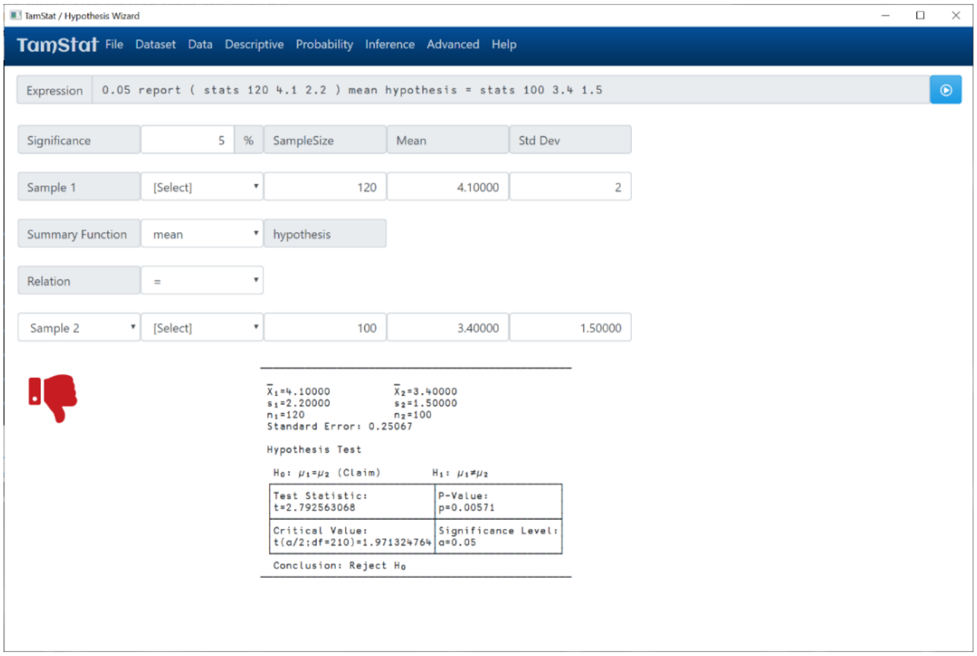
Expression `0.05 report (State in 'PA') proportion hypothesis = 0.3`

Significance	0.05	Sample Size	Proportion	Events
Sample	State in 'I	38	0.4210526316	16
Summary Function	proportion	hypothesis		
Relation	=			
Hypothesized Va	0.3			

```
^
p =0.42105
n =38
Standard Error: 0.07434

Hypothesis Test

 H0: p=0.3 (Claim)     H1: p≠0.3
 Test Statistic:        P-Value:
 Z=1.628382348          p=0.10344
 Critical Value:        Significance Level:
 Z(α/2)=1.959962728     α=0.05
 Conclusion: Fail to reject H0
```

A supermarket is testing two checkout lane designs. For System A, 120 customer checkout times produced a sample mean of 4.1 and a sample standard deviation of 2.2 minutes. For System B, 100 customer checkout times produced a sample mean of 3.4 and a sample standard deviation of 1.5 minutes. Test at the 1% significance level, the hypothesis that the two systems are the same.

To perform a two-sample test, change "Hypothesized Value" to "Sample 2" by selecting the drop-down:


TamStat / Hypothesis Wizard
TamStat File Dataset Data Descriptive Probability Inference Advanced Help

Expression `0.05 report ( stats 120 4.1 2.2 ) mean hypothesis = stats 100 3.4 1.5`

Significance	5 %	SampleSize	Mean	Std Dev
Sample 1	[Select]	120	4.10000	2
Summary Function	mean	hypothesis		
Relation	=			
Sample 2	[Select]	100	3.40000	1.50000

```
X̄1=4.10000          X̄2=3.40000
s1=2.20000          s2=1.50000
n1=120              n2=100
Standard Error: 0.25067

Hypothesis Test

 H0: μ1=μ2 (Claim)          H1: μ1≠μ2
 Test Statistic:            P-Value:
 t=2.792563068              p=0.00571
 Critical Value:            Significance Level:
 t(α/2;df=210)=1.971324764  α=0.05
 Conclusion: Reject H0
```

3.7 Exercises

1. From the Student data set estimate the proportion of students from Pennsylvania using 80%, 90%, 95% and 99% confidence intervals. What sample size would be required if you wanted the margin of error for a 95% confidence interval to be less than 2%?
2. You are conducting a political poll and desire that your results be accurate to 4 percentage points with 90% confidence. What is the minimum sample size required?
3. A previous poll revealed that a state governor had a 35% favorable performance rating last year. What sample size should be used to conduct a poll next year to be accurate to within 3 percentage points with 95% confidence? If 700 people responded favorably, find a 95% confidence interval.

4. For the Student data, calculate a 99% confidence interval for the mean weight. What sample size do you need for the Margin of Error to be 5 lbs.?

5. For the Student data, calculate a 99% confidence interval for the mean number of siblings. What sample size do you need for the Margin of Error to be 0.2?

6. An English teacher wants to determine whether the mean reading speed of a certain student is at least 600 words per minute. What can he conclude if in six 1-minute intervals, the student reads 606, 622, 617, 572, 570, and 605 words and the probability of a type I error is to be at most 0.05?
7. It has been claimed that more than 70% of the students attending a large university are opposed to a plan to increase student fees in order to build new parking facilities. If 150 of 180 students selected at random at that university are opposed to the plan, test the claim at the 0.05 level of significance.
8. A company claims that the variance of the sugar content of its yogurt is less than or equal to 25. (The sugar content is measured in milligrams per ounce.) A sample of 20 servings is selected and the sugar content is measured. The variance of the sample is found to be 36. At α=0,10, is there enough evidence to reject the claim?
9. A study of 36 members of the central park walkers showed that they could walk at an average of 2.6 mph. The sample standard deviation is 0.4. Find a 95% confidence interval.
10. In a survey of 1004 individuals, 442 felt that the president spent too much time away from Washington. Find a 95% confidence interval for the true population proportion.
11. An insurance company is trying to estimate the average number of sick days that full-time food service workers use per year. A pilot study found the standard deviation to be 2.5 days. How large a sample must be selected if the company wants to be 95% confident of getting an interval that contains the true mean with a maximum error of 1 day?
12. How large a sample should be surveyed to estimate the true proportion of college students who do laundry once a week within 3% with 95% confidence?

Chapter 4 Analysis of Variance

Sometimes comparisons can take place between two or more groups (or levels). Comparisons can also take place using one or more factors. We will first look at One-Way ANOVA where we compare three or more groups and at other designs when there are more than one factor. These include blocked designs, Latin-Square designs, two-and three factor designs with fixed factors, random factors and mixed factors, factorial designs, then complete and fractional 2^k factorial designs, and finally nested designs.

4.1 One-Way ANOVA

We have shown how to determine if means differ for two populations. Suppose we wish examine three or more populations. Our null hypothesis for three populations would be:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

The alternative hypothesis would be that at least one of the populations has a different mean. It would not be:

$H_1: \mu_1 \neq \mu_2 \neq \mu_3$, rather it would be: $H_1: (\mu_1 \neq \mu_2) or (\mu_2 \neq \mu_3)$
In general the null hypothesis for three or more populations would be: $H_0: (\forall i \neq j)\ \mu_i = \mu_j$

And the alternative hypothesis would be: $H_1: (\exists i \neq j)\ \mu_i \neq \mu_j$ (The symbols ∀ and ∃ mean "for all" and "there exists" respectively).

A pharmaceutical company wants to investigate the effectiveness of a new drug. A single-factor experiment was conducted with three dosage levels: 20, 30 and 40 grams. Four replications were applied to each level.

```
    G20←24 28 37 30  ⍝ Dosage = 20 grams; n1=4
    G30←37 44 31 35  ⍝ Dosage = 30 grams; n2=4
    G40←42 47 52 38  ⍝ Dosage = 40 grams; n3=4

    NS←oneWay anova G20 G30 G40

    NS.AnovaTable
 SOURCE           SS  DF            MS           F             P
 A       450.6666667   2  225.3333333   7.03555941  0.01446379777
 Error   288.25        9   32.02777778  0           0
 Total   738.9166667  11    0           0           0

     NS.Means
 Group  N  Mean      StdDev
 1      4 29.75  5.439056291
 2      4 36.75  5.439056291
 3      4 44.75  6.075908711
```

If you want to label the levels, you can list the namespace followed by the variable names in quotes:

```
      NS← oneWay anova ## 'G20 G30 G40'
      NS.Means
 Group  N  Mean      StdDev
 G20    4 29.75  5.439056291
 G30    4 36.75  5.439056291
 G40    4 44.75  6.075908711
```

A printable report consists of the ANOVA table with R-Squared data, group means and confidence intervals, and a comparison between treatment means using the least significant difference method.

```
    report  NS
```

```
 ANOVA Table

 SOURCE              SS    DF              MS         F         P  F(α=.05)
 ------ --------------- ----- --------------- --------- --------- ---------
 A               450.67     2          225.33      7.04   0.01446      4.26
 Error           288.25     9           32.03
 ------ --------------- ----- --------------- --------- ---------
 Total           738.92    11

 S =    5.65931  R-Sq =  60.99%  R-Sq(adj) =   52.32%

 Group  N  Mean       StdDev
 G20    4 29.75  5.439056291
 G30    4 36.75  5.439056291
 G40    4 44.75  6.075908711

Confidence Interval

95% G20(23.349,36.151)       (------------*-----------)
95% G30(30.349,43.151)                     (------------*-----------)
95% G40(38.349,51.151)                                     (------------*-----------)
                      +---------+---------+---------+---------+---------+---------+-
                               25        30        35        40        45        50

 Treatment   Mean Difference  DF       SE         T      P    LSD
  G20 vs G30              ¯7   1  4.0017  ¯1.7492 0.1142 9.0526
  G20 vs G40             ¯15   1  4.0017  ¯3.7484 0.0046 9.0526
  G30 vs G40              ¯8   1  4.0017  ¯1.9991 0.0767 9.0526
```

One of the variables in the student survey was "State". One would suspect that the heights of students would not vary by home state. To test this hypothesis, we will use the variable "Height" and partition it using the variable "State". Note that the sample sizes for the five states are different.

```
     frequency #.SD.State
 CT   4
 MD   3
 NJ   6
 NY   9
 PA  16
```

Since we require a minimum of three groups to perform a one-way ANOVA, if there are only two items in the right argument, the first is the variable of interest, and the second is a grouping variable. Both must have the same length.

```
      NS←oneWay anova #.SD.Height #.SD.State
```

An alternative way is for the right argument to consist of a namespace followed by a name list of variables within that namespace. In the one-way anova, this would consist of the name of the variable of interest and the name of the grouping variable. In our case, we would enter:

```
      NS←oneWay anova #.SD 'Height State'
```

We could then generate a printable report using the report function. For a significance level $\alpha \neq 0.05$ we can provide a left argument to **report**:

```
    .10 report NS
```

```
 ANOVA Table

 SOURCE              SS    DF              MS         F         P  F(α=.10)
 ------ --------------- ----- --------------- --------- --------- ---------
 A                63.11     4           15.78      0.77   0.55401      2.12
 Error           678.24    33           20.55
 ------ --------------- ----- --------------- --------- ---------
 Total           741.35    37

 S =    4.53352  R-Sq =   8.51%  R-Sq(adj) =   ¯2.58%

 Group   N       Mean        StdDev
  PA    16 68.3125      4.331570154
  MD     3 71.33333333  2.516611478
  CT     4 71.25        0.9574271078
  NJ     6 67.16666667  6.554896389
  NY     9 68.72222222  4.562832941

Confidence Interval

90%  PA (66.394,70.231)                                (---------*--------)
90%  MD (66.904,75.763)                                   (---------------------*---------------------)
90%  CT (67.414,75.086)                                     (------------------*------------------)
90%  NJ (64.034,70.299)                    (---------------*--------------)
90%  NY (66.165,71.280)                               (------------*-----------)
                       +---------+---------+---------+---------+---------+---------+---------+---------+------
                                62        64        66        68        70        72        74        76

Treatment     Mean Difference  DF      SE         T      P    LSD
  PA  vs  MD            ¯3.0208   1   2.8523  ¯1.0591 0.2972 4.8271
  PA  vs  CT            ¯2.9375   1   2.5343  ¯1.1591 0.2547 4.289
  MD  vs  CT             0.0833   1   3.4625   0.0241 0.9809 5.8599
  PA  vs  NJ             1.1458   1   2.1703   0.528  0.6011 3.6729
  MD  vs  NJ             4.1667   1   3.2057   1.2998 0.2027 5.4252
  CT  vs  NJ             4.0833   1   2.9264   1.3954 0.1722 4.9525
  PA  vs  NY            ¯0.4097   1   1.889   ¯0.2169 0.8296 3.1968
  MD  vs  NY             2.6111   1   3.0223   0.8639 0.3939 5.1149
  CT  vs  NY             2.5278   1   2.7243   0.9279 0.3602 4.6105
  NJ  vs  NY            ¯1.5556   1   2.3894  ¯0.651  0.5195 4.0437
```

Note that the least significant difference (LSD) values are different because the sample sizes are different. In the previous example, they were all the same because the sample sizes were identical.

4.2 The Randomized Complete Block Design

The randomized complete block design is an extension of the paired t-test to more than two groups. We want to test whether four different tips produce different hardness readings on a machine. The tips are pressed into a metal specimen and the depth of the impression is measured. There is also some variation in the metal specimens. To remove this excess variation, we test each of the four tips on each of four specimens. The measurements are as follows:

Tip	Specimen 1	Specimen 2	Specimen 3	Specimen 4
1	9.3	9.4	9.6	10.0
2	9.4	9.3	9.8	9.9
3	9.2	9.4	9.5	9.7
4	9.7	9.6	10.0	10.2

The right argument for a randomized complete blocked design is a numeric matrix; the rows represent treatments and the columns represent blocks.

```
      X←4 4⍴9.3 9.4 9.6 10 9.4 9.3 9.8 9.9 9.2 9.4 9.5 9.7 9.7 9.6 10 10.2
9.3 9.4  9.6 10
9.4 9.3  9.8  9.9
9.2 9.4  9.5  9.7
9.7 9.6 10   10.2
```

To generate a blocked design, we use the **blocked** operand:

```
     report blocked anova X
────────────────────────────────────────────────────────────────────────────────
 ANOVA Table

 SOURCE                  SS    DF              MS         F         P  F(α=.05)
 ------     --------------- ----- --------------- --------- --------- ---------
 Treatments          0.3850     3          0.1283     14.44   0.00087      3.86
 Blocks              0.8250     3          0.2750
 Error               0.0800     9          0.0089
 ---------- --------------- ----- --------------- --------- ---------
 Total               1.2900    15

 S =    0.09428  R-Sq =  93.80%  R-Sq(adj) =   89.66%
────────────────────────────────────────────────────────────────────────────────
```

4.3 The Latin Square Design

When there are two extraneous sources of variation, we can using a special blocking design known as the Latin Square Design. A Latin Square contains the same number of rows and columns, where each cell contains a letter designation, where each letter occurs once and only once in each row and column. A cyclic Latin Square shifts each row by one letter. An example of a 4 x 4 cyclic Latin square is:

```
        LS←0 1 2 3 4⌽5 5⍴'ABCDE'

BCDEA
CDEAB
DEABC
EABCD
```

The letters in a Latin Square represent the Treatments; the rows and columns are the blocks. To use a Latin Square design in TamStat, set up the right argument as an N x N square matrix of measurements. The left argument is a simple character matrix representing the Latin Square. If the left argument is omitted, the default is a cyclic Latin Square of the same shape as the right argument.

Five different formulations of an explosive mixture used in dynamite are being studied. We label these A, B, C, D and E. There are 5 batches of raw material and 5 operators. 25 experiments are conducted; each row represents a batch of raw material, while each column represents an operator. The results of the experiment are:

```
     X←5 5⍴24 20 19 24 24 17 24 30 27 36 18 38 26 27 21 26 31 26 23 22 22 30 20 29 31

      LS{α '=' ω}¨X
 A= 24  B= 20  C= 19  D= 24  E= 24
 B= 17  C= 24  D= 30  E= 27  A= 36
 C= 18  D= 38  E= 26  A= 27  B= 21
 D= 26  E= 31  A= 26  B= 23  C= 22
 E= 22  A= 30  B= 20  C= 29  D= 31
```

To generate an ANOVA table, use the `latinSquare` function:

```
    latinSquare X
Treatments  330  4 82.5        7.734375 0.02974249554
Rows         68  4 17          0        0
Columns     150  4 37.5        0        0
Error       128 12 10.66666667 0        0
```

To get a printable report, enter the following:

```
report latinSquare anova X
ANOVA Table

SOURCE                  SS    DF              MS         F         P  F(α=.05)
------     --------------- ----- --------------- --------- --------- ---------
Treatments          330.00     4           82.50      7.73   0.00254      3.26
Rows                 68.00     4           17.00
Columns             150.00     4           37.50
Error               128.00    12           10.67
---------- --------------- ----- --------------- --------- ---------
Total               676.00    24

S =    3.26599  R-Sq =  81.07%  R-Sq(adj) =   62.13%
```

4.4 Factorial Designs

When the effects of more than one factor are being studied, we can use a factorial design. Factorial designs are most effective when they are balanced, i.e. each treatment combination has the same number of replications. TamStat currently handles one, two and three-factor factorial designs. One-factor designs are equivalent to using the `oneWay` operand except that the sample size for each level must be the same. The right argument can be a matrix or higher-rank array. The last dimension(s) are assumed to be replications. The left argument is 1, 2 or 3, depending upon the number of factors. Normally the rank of the right argument is one more that the left argument. If the rank equals the right argument, then a factorial design with a single replicate is performed. If the rank of the right argument minus the left argument is two or more, the trailing axes will be catenated together.

Material type and temperature can affect the maximum output voltage of a battery. An experiment is conducted with three materials, three temperatures (50, 60 and 80 degrees F) ,and four replications. The following are the results of the experiment:

```
   X←130 155 74 180,34 40 80 75, 20 70 82 58
   X,←150 188 159 126, 136 122 106 115, 25 70 58 45
   X,←138 110 168 160, 174 120 150 139, 96 104 82 60
   X←3 3 4⍴X

   ('Material' 'Type1' 'Type2' 'Type3'),'50 deg F' '65 deg F' '80 deg F'⍪↓X

Material  50 deg F         65 deg F         80 deg F
Type1     130 155 74 180   34 40 80 75      20 70 82 58
Type2     150 188 159 126  136 122 106 115  25 70 58 45
Type3     138 110 168 160  174 120 150 139  96 104 82 60
```

```
     report 2 factor anova X
```

```
 ANOVA Table

 SOURCE                SS    DF               MS         F         P  F(α=.05)
 ------ --------------- ----- --------------- --------- --------- ---------
 A                 10,684     2            5,342      7.91   0.00198      3.35
 B                 39,119     2           19,559     28.97  <0.00001      3.35
 AB                 9,614     4            2,403      3.56   0.01861      2.73
 Error             18,231    27              675
 ------ --------------- ----- --------------- --------- ---------
 Total             77,647    35

 S =   25.98486  R-Sq =  76.52%  R-Sq(adj) =   69.56%
                                                                                2
```

If you wish to label the factors, you can replace the “2” with a 2-item vector of factor names:

```
 ANOVA Table

 SOURCE                       SS    DF              MS         F         P  F(α=.05)
 ------       --------------- ----- --------------- --------- --------- ---------
 Material                 10,684     2           5,342      7.91   0.00198      3.35
 Temp                     39,119     2          19,559     28.97  <0.00001      3.35
 MaterialTemp              9,614     4           2,403      3.56   0.01861      2.73
 Error                    18,231    27             675
 ------------ --------------- ----- --------------- --------- ---------
 Total                    77,647    35

 S =   25.98486  R-Sq =  76.52%  R-Sq(adj) =   69.56%
```

A soft drink bottler wants to obtain more uniform fill heights; the factors of interest are (A) % carbonation, (B) operating pressure, and (C) line speed. Carbonation levels are 10, 12 and 14%. Pressure is at two levels: 25 and 30 psi. And line speeds are 200 and 250 bottles per minute. The fill heights are displayed below in row major order (negative heights are below target, positive heights are above target.)

```
     Y←¯3 ¯1 ¯1 0 ¯1 0 1 1 0 1 2 1 2 3 6 5 5 4 7 6 7 9 10 11
     report 3 factor anova 3 2 2 2⍴Y
```

```
 ANOVA Table
 SOURCE                SS    DF               MS         F         P  F(α=.05)
 ------ --------------- ----- --------------- --------- --------- ---------
 A                 252.75     2           126.38    178.41  <0.00001      3.89
 B                  45.38     1            45.38     64.06  <0.00001      4.75
 C                  22.04     1            22.04     31.12   0.00012      4.75
 AB                  5.25     2             2.63      3.71   0.05581      3.89
 AC                  0.58     2             0.29      0.41   0.67149      3.89
 BC                  1.04     1             1.04      1.47   0.24859      4.75
 ABC                 1.08     2             0.54      0.76   0.48687      3.89
 Error               8.50    12             0.71
 ------ --------------- ----- --------------- --------- ---------
 Total             336.63    23

 S =    0.84163  R-Sq =  97.47%  R-Sq(adj) =   95.16%
```

4.5 Random and Mixed Models

When one or more of the factors are random, the calculation of the test statistics (F values) may not always use the mean square error (MSE). In the random model, both factors are random. In the mixed model, the first factor (A) is fixed, and the second factor (B) is random. In both these models, we can still use the factor operand; however, the left argument is now a two-item vector consisting of the number of fixed factors followed by the number of random factors.

Using the battery voltage in the previous example, assume that material types and temperatures were randomly selected from a larger population. Since both factors are random, the left argument is 0 2 indicating no fixed factors and 2 random factors:

```
     report 0 2 factor anova X
```

```
ANOVA Table

SOURCE              SS    DF              MS         F         P  F(α=.05)
------ --------------- ----- --------------- --------- --------- ---------
A               10,684     2           5,342      2.22   0.22434      3.35
B               39,119     2          19,559      8.14   0.03892      3.35
AB               9,614     4           2,403      3.56   0.01861      2.73
Error           18,231    27             675
------ --------------- ----- --------------- --------- ---------
Total           77,647    35

S =   25.98486  R-Sq =  76.52%  R-Sq(adj) =   69.56%
```

Note that the F statistics for the main effects are much different than in the fixed model. In particular. the statistics $F_A = MS_A/MS_{AB}$ and $F_B = MS_B/MS_{AB}$.

Sometimes, one factor is fixed and the other is random. In the previous example suppose that the materials (A) are fixed, but the temperatures (B) are random. The left argument would be 1 1: one fixed factor followed by one random factor:

```
     report 1 1 factor anova X
```

```
ANOVA Table

SOURCE              SS    DF              MS         F         P  F(α=.05)
------ --------------- ----- --------------- --------- --------- ---------
A               10,684     2           5,342      2.22   0.22434      3.35
B               39,119     2          19,559     28.97  <0.00001      3.35
AB               9,614     4           2,403      3.56   0.01861      2.73
Error           18,231    27             675
------ --------------- ----- --------------- --------- ---------
Total           77,647    35

S =   25.98486  R-Sq =  76.52%  R-Sq(adj) =   69.56%
```

In this case, the statistics $F_A = MS_A/MS_{AB}$ and $F_B = MS_B/MS_E$.

4.6 2*K Factorial Designs

When each factor has exactly two levels, there are no limits to the number of factors in TamStat. The sum of squares calculations in 2^k Factorial designs are much simpler.

Instead of creating a k-dimension array as required by the `factor` operand, the `factorial2k` operand allows us to use a vector of shape `2*K` or a matrix of shape `(2*K),R` where R is the number of replications. Each item or row of the vector or matrix respectively corresponds to a particular treatment combination. The treatment combinations must be in standard order. Since there are only two levels for each factor – high(+) and low(-), each treatment combination can be described with a list of lower-case letters representing which factors are run at the higher level. For example (1) means all factors are set to the low level (-); *a* means that factor A is set to the high level and all the others are at the low level, ac means that factors A and C are at the high level and factor B (and any others) are set at the low level. For a `2*3` factorial design, standard order is:

(1), a, b, ab, c, ac, ab, abc

The TamStat function standardOrder generates a list of treatment combinations in standard order. The right argument is a list of factors set to the high (+) level.

```
      standardOrder 'ab'         ⍝ Two factors
(1)  a  b  ab
      standardOrder 'abc'        ⍝ Three factors
(1)  a  b  ab  c  ac  bc  abc
     standardOrder 'abcd'        ⍝ Four factors
(1)  a  b  ab  c  ac  bc  abc  d  ad  bd  abd  cd  acd  bcd  abcd
```

Note that the first four items of three factors and four factors are the same as two factors.

Effects are indicated with capital letters, e.g. `A` is the main effect for factor A, `AB` is the interaction effect. `I` represents the sum of all treatment combinations. To show the relationship between treatment combinations and effects for a 3-factor factorial design, use the **`hadamard`** function.

```
    hadamard show 3
    I A  B  AB  C  AC  BC  ABC
(1) + -  -  +   -  +   +   -
a   + +  -  -   -  -   +   +
b   + -  +  -   -  +   -   +
ab  + +  +  +   -  -   -   -
c   + -  -  +   +  -   -   +
ac  + +  -  -   +  +   -   -
bc  + -  +  -   +  -   +   -
abc + +  +  +   +  +   +   +
```

In the fill volume example in the previous section, suppose only two levels are carbonation 10% and 12% are used. This would make this a 2*3 factorial design. Arranging the data in standard order (in columns to save space):

Replicate	(1)	a	b	ab	c	ac	bc	abc
1	-3	0	-1	2	-1	2	1	6
2	-1	1	0	3	0	1	1	5

¶

```
    X←8 2⍴¯3 ¯1 0 1 ¯1 0 2 3 ¯1 0 2 1 1 1 6 5 ⍝ Rows = treatment combinations
```

```
     report   factorial2k anova X
```

```
 ANOVA Table

 SOURCE                SS    DF               MS         F         P  F(α=.05)
 ------ --------------- ----- --------------- --------- --------- ---------
 A                 36.000     1           36.000     57.60   0.00006      5.32
 B                 20.250     1           20.250     32.40   0.00046      5.32
 C                 12.250     1           12.250     19.60   0.00221      5.32
 AB                 2.250     1            2.250      3.60   0.09435      5.32
 AC                 0.250     1            0.250      0.40   0.54474      5.32
 BC                 1.000     1            1.000      1.60   0.24150      5.32
 ABC                1.000     1            1.000      1.60   0.24150      5.32
 Error              5.000     8            0.625
 ------ --------------- ----- --------------- --------- ---------
 Total             78.000    15
```

Here we see that the three main effects are significant, but none of the interactions are.

4.6.1 A Single Replication of 2*K Design

When there are a large number of factors, the number of replications is limited and, in many cases, only a single replication is necessary. A factorial experiment is conducted to study the factors affecting the filtration rate of a chemical product. The four factors are (A) temperature (B) pressure, (C) reactant concentration, and (D) stir rate The results of the single-replicate experiment are shown in the following table:

Filtration Data from Experiment		C0			C1	
		D0	D1		D0	D1
A0	B0	45	43		68	75
	B1	48	45		80	70
A1	B0	71	100		60	86
	B1	65	104		65	96

In the above chart A0 is the lower level and A1 is the upper level of factor A. For example, the treatment combination $ac = 60$ and the treatment combination $d = 43$. When there is a single replicate, we use a vector of length `2*K.` Since standard order is not the same as row major order used in factorial designs, it can be tedious to construct a vector in standard order from the above table. We would start with (1), a, b ab, … Note that $(1) = 45$ $a = 71$, $b = 48$, and $ab = 65$. These values are scattered in the above table. Since there is a single replicate, we can enter the data in vector format. Let's enter them directly from the table in row major order

```
      FiltrationRate←45 43 68 75 48 45 80 70 71 100 60 86 65 104 65 96
```

We can easily convert the data to standard order using the following utility. Note that the first four items match the treatment combinations (1), a, b, and b respectively:

```
      X←rowmajor2std FiltrationRate
45 71 48 65 68 60 80 65 43 100 45 104 75 86 70 96
```

The following shows all the treatment combinations and the results of each in standard order:

```
      ↑(standardOrder 'abcd')X
 (1)   a   b  ab   c  ac  bc  abc   d   ad  bd  abd  cd  acd  bcd  abcd
  45  71  48  65  68  60  80   65  43  100  45  104  75   86   70    96
```

We are now ready to perform the analysis of variance:

```
    report factorial2k anova X
```

```
ANOVA Table

SOURCE              SS    DF              MS         F         P  F(α=.05)
------ -------------- ----- -------------- --------- --------- ---------
A             1,870.6     1        1,870.6     73.18   0.00036      6.61
B                39.1     1           39.1      1.53   0.27130      6.61
C               390.1     1          390.1     15.26   0.01134      6.61
D               855.6     1          855.6     33.47   0.00217      6.61
AB                0.1     1            0.1             0.96248      6.61
AC            1,314.1     1        1,314.1     51.41   0.00082      6.61
AD            1,105.6     1        1,105.6     43.25   0.00122      6.61
BC               22.6     1           22.6      0.88   0.39061      6.61
BD                0.6     1            0.6      0.02   0.88787      6.61
CD                5.1     1            5.1      0.20   0.67491      6.61
Error           127.8     5           25.6
------ -------------- ----- -------------- --------- ---------
Total         5,730.9    15

S =    5.05594  R-Sq =  97.77%  R-Sq(adj) =   93.31%
```

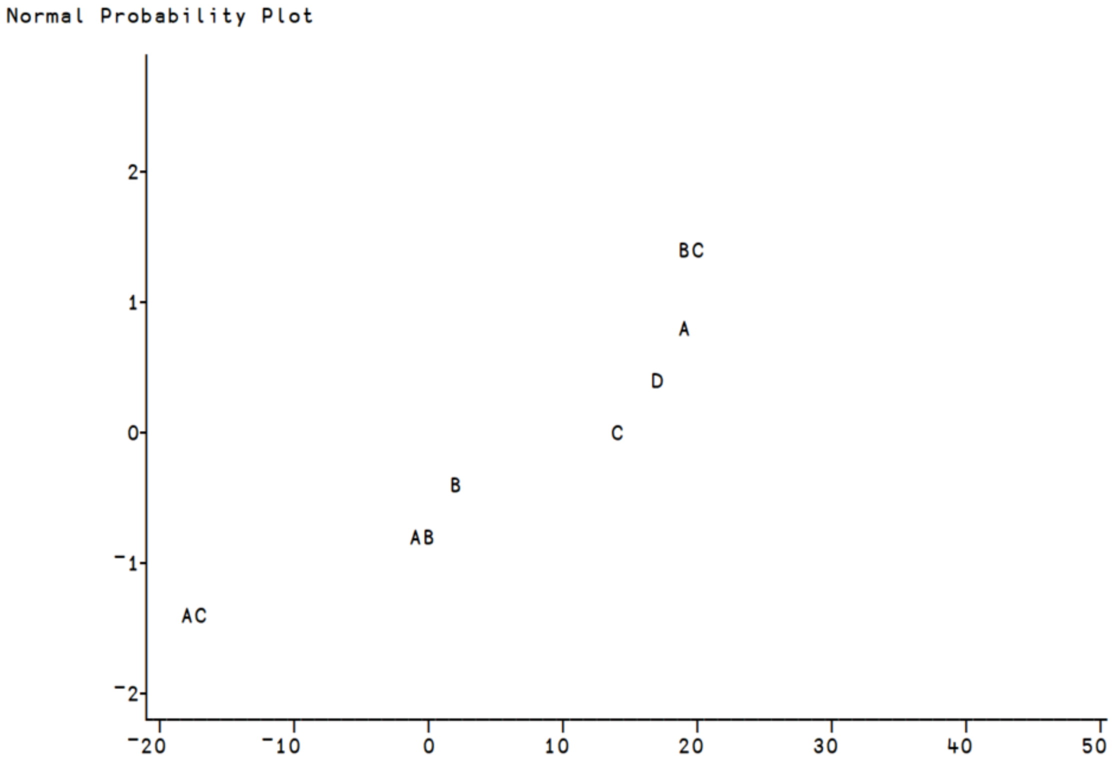
The results show that main effects A, C and D and the interactions AC and AD are significant.

4.6.2 Fractional Factorial Designs

A fractional factorial design allows one to ignore high-order interactions and concentrate on main effects and two-factor interactions. Main effects are confounded with high order interactions resulting in running fewer experiments, e.g. a one-half fractional factorial design requires half the experiments as a full factorial. To accomplish this, we need to create an alias structure. The alias is determined by generators which equal the identity element I. To analyze a fractional factorial design in TamStat, simply provide the generator) as the optional left argument. In the previous example, we had 16 runs. A one-half fractional factorial design will let us use only 8 runs. We will use `I=ABCD` as the defining relation, so the generator is simply ABCD.

```
     report   'ABCD' factorial2k anova 45 100 45 65 75 60 80 96
```

```
Fractional Factorial Design

 Variable  Alias  Contrast  Effect     SS     Pct
 [A]       A+BCD        76    19     722    23.506
 [B]       B+ACD         6     1.5     4.5   0.147
 [C]       C+ABD        56    14     392    12.762
 [D]       D+ABC        66    16.5   544.5  17.727
 [AB]      AB+CD        ¯4    ¯1       2     0.065
 [AC]      AC+BD       ¯74   ¯18.5   684.5  22.286
 [BC]      BC+AD        76    19     722    23.506
```

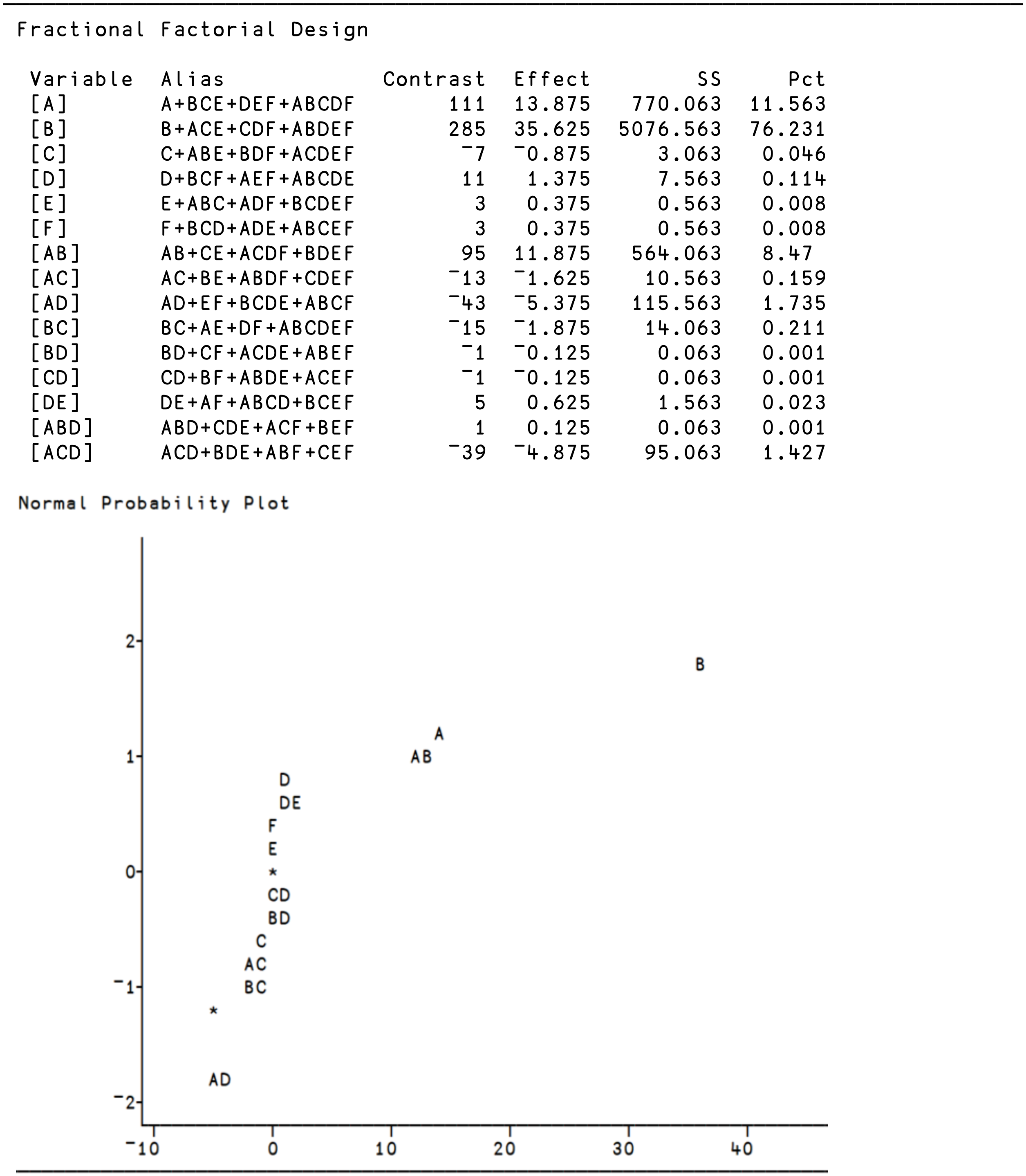


Notice that the TamStat output is somewhat different. Fractional factorial designs usually contain a single replicate so it is not possible to generate a traditional ANOVA table until we can determine which effects are significant and which are not. The normal probability plot may give some clues as to which effects are significant.

Parts manufactured in an injection molding process show excessive shrinkage, which causes problems in assembly operations downstream. A designed experiment to reduce shrinkage involves 6 factors: mold temperature A, screw speed B, holding time C, cycle time D, gate size E, and holding pressure F—each at two levels:

Run	A	B	C	D	E=ABC	F=BCD	Shrinkage x 10	
1	-	-	-	-	-	-	6	(1)
2	+	-	-	-	+	-	10	ae
3	-	+	-	+	+	+	21	bef
4	+	+	-	-	-	+	60	abf
5	-	-	+	-	+	+	4	cef
	+	-	+	-	-	+	15	acf
7	-	+	+	-	-	-	26	bc
8	+	+	+	-	+	-	60	abce
9	-	-	-	+	-	+	8	df
10	+	-	-	+	+	+	12	adef
11	-	+	-	+	+	-	34	bde
12	+	+	-	+	-	-	60	abd
13	-	-	+	+	+	-	16	cde
14	+	-	+	+	-	-	5	acd
15	-	+	+	+	-	+	37	bcdf
16	+	+	+	+	+	+	52	abcdef

```
X ← 6 10 32 60 4 15 26 60 8 12 34 60 16 5 37 52
report 'ABCE=BCDF' factorial2k anova X
```



```
Fractional Factorial Design

 Variable  Alias              Contrast  Effect         SS      Pct
 [A]       A+BCE+DEF+ABCDF         111  13.875    770.063   11.563
 [B]       B+ACE+CDF+ABDEF         285  35.625   5076.563   76.231
 [C]       C+ABE+BDF+ACDEF          ¯7  ¯0.875      3.063    0.046
 [D]       D+BCF+AEF+ABCDE          11   1.375      7.563    0.114
 [E]       E+ABC+ADF+BCDEF           3   0.375      0.563    0.008
 [F]       F+BCD+ADE+ABCEF           3   0.375      0.563    0.008
 [AB]      AB+CE+ACDF+BDEF          95  11.875    564.063    8.47
 [AC]      AC+BE+ABDF+CDEF         ¯13  ¯1.625     10.563    0.159
 [AD]      AD+EF+BCDE+ABCF         ¯43  ¯5.375    115.563    1.735
 [BC]      BC+AE+DF+ABCDEF         ¯15  ¯1.875     14.063    0.211
 [BD]      BD+CF+ACDE+ABEF          ¯1  ¯0.125      0.063    0.001
 [CD]      CD+BF+ABDE+ACEF          ¯1  ¯0.125      0.063    0.001
 [DE]      DE+AF+ABCD+BCEF           5   0.625      1.563    0.023
 [ABD]     ABD+CDE+ACF+BEF           1   0.125      0.063    0.001
 [ACD]     ACD+BDE+ABF+CEF         ¯39  ¯4.875     95.063    1.427

Normal Probability Plot
```

4.7 Nested Designs

In nested designs, the levels of one factor are dependent upon another factor. For example. A company buys supplies from three different suppliers. Four batches of raw material are randomly selected from each supplier, and three measurements(replications) are taken on each batch. The Suppliers represent factorA. The batches represent factor B; however, there is no relationship between the batches for the first supplier and the other suppliers. The data are listed below:

Supplier	1				2				3			
Batches	1	2	3	4	5	6	7	8	9	10	11	12
Data	94	91	91	94	94	93	92	93	95	91	94	96
	92	90	93	97	91	97	93	96	97	93	92	95
	93	89	94	93	90	95	91	95	93	95	95	94

```
    X2←94 92 93 91 90 89 91 93 94 94 97 93   ⍝ Supplier 1
    X2,←94 91 90 93 97 95 92 93 91 93 96 95   ⍝ Supplier 2
    X2,←95 97 93 91 93 95 94 92 95 96 95 94   ⍝ Supplier 3

       report nested anova 3 4 3⍴X2
```

```
ANOVA Table

SOURCE              SS    DF              MS         F         P  F(α=.05)
------ --------------- ----- --------------- --------- --------- ---------
A                15.06     2            7.53      0.97   0.41578      3.40
B(A)             69.92     9            7.77      2.94   0.01667      2.30
Error            63.33    24            2.64
------ --------------- ----- --------------- --------- ---------
Total           148.31    35

S =    1.62447  R-Sq =  57.30%  R-Sq(adj) =   37.72%
```

Notice that since B is nested within A we use the notation B(A) to represent the batches within supplier.

In the next example, we have three factors, and the third factor is nested within the second factor.

An industrial engineer is trying to increase the assembly speed of electronic components. He is looking at three assembly fixes and two workplace layouts. Since the workplaces are in different locations, he cannot use the same operators, so operators are nested within workplace layouts. There are four operators assigned to each treatment combination.

	Layout 1				Layout 2			
Operator	1	2	3	4	5	6	7	8
Fixture 1	22	23	28	25	26	27	28	24
	24	24	29	23	28	25	25	23
Fixture 2	30	29	30	27	29	30	24	28
	27	28	32	25	28	27	23	30
Fixture 3	25	24	27	26	27	26	24	28
	21	22	25	23	25	24	27	27

```
    X3←22 24 23 24 28 29 25 23 26 28 27 25 28 25 24 23
    X3,←30 27 29 28 30 32 27 25 29 28 30 27 24 23 28 30
    X3,←25 21 24 22 27 25 26 23 27 25 26 24 24 27 28 27
```

```
              report 'FLO' nested anova 3 2 4 2ρX3

ANOVA Table

SOURCE              SS    DF              MS         F         P  F(α=.05)
------ --------------- ----- --------------- --------- --------- ---------
F                82.79     2           41.40      7.55   0.00755      3.40
L                 4.08     1            4.08      0.34   0.58070      4.26
O(L)             71.92     6           11.99      5.14   0.00161      2.51
FL               19.04     2            9.52      1.74   0.21777      3.40
FO(L)            65.83    12            5.49      2.35   0.03604      2.18
Error            56.00    24            2.33
------ --------------- ----- --------------- --------- ---------
Total           299.67    47

S =    1.52753  R-Sq =  81.31%  R-Sq(adj) =   63.40%
```

Observe that the left argument to nested can override the standard A, B, C levels. We are using F to represent Fixture, L to represent Layout and O to represent Operator.

4.8 Summary Table

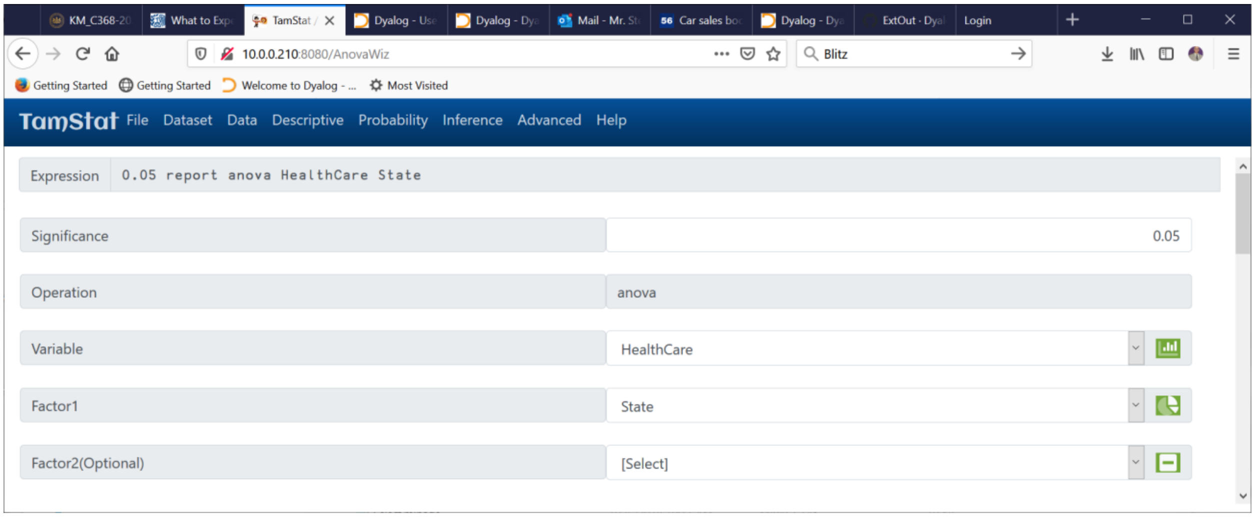
The table on the next page shows all the possible operands to the `anova` operator, the optional left arguments, the right argument format and the results.

The TamStat `anova` Operator

Design	Operand	Left Argument	Right Argument	Result
One Factor; 3 or more groups; sample sizes may vary	`oneWay`	None	- Vector of numeric vectors - 2-item Vector of database variables - NameSpace NameList	- Anova Table - R-Squared/S - Group Means - Confidence Intervals - LSD Analysis
Blocked Replications	`blocked`	None	- Numeric Matrix, blocks are columns - Numeric Array; blocks in last dimension	- Anova Table - R-Squared/S
Treatments and 2 blocks	`latinSquare`	Square Character Matrix (Optional)	- Square Numeric Matrix (n x n)	- Anova Table - R-Squared/S
Balanced Factorial Design; 1, 2 or 3 factors	`factor`	Numeric Scalar: 1, 2 or 3	- Numeric Matrix - Numeric Array	- Anova Table - R-Squared/S - If left arg=1, same as oneWay
Random Factors (2 random factors)	`factor`	Numeric Vector: `0 2`	- Numeric Array	- Anova Table - R-Squared/S
Mixed Factors (One Fixed, one random)	`factor`	Numeric Vector: `1 1`	- Numeric Array	- Anova Table - R-Squared/S
2^k Factorial	`factorial2k`	None	- Vector of length `2*K` - Matrix `(2*K),R`	- Anova Table - R-Squared/S
Fractional 2^k Factorial	`factorial2k`	Alias(es) - Character Vector	- Vector of length `2*K` - Matrix `(2*K),R`	- Anova Table - R-Squared/S
Nested Design; 2 or 3 factors; last factor nested within previous.	`nested`	Factor Names (Optional)	- Numeric Array	- Anova Table - R-Squared/S

4.9 ANOVA Wizard

From the main menu, select “Advanced Topics” then “ANOVA”.   To perform a one-way ANOVA, select a quantitative variable for “Variable” and a qualitative variable for “Treatment”. For example, we will select the numerical response to the Healthcare variable, and the variable STATE as the qualitative variable:



Expression `0.05 report anova HealthCare State`

Significance: 0.05

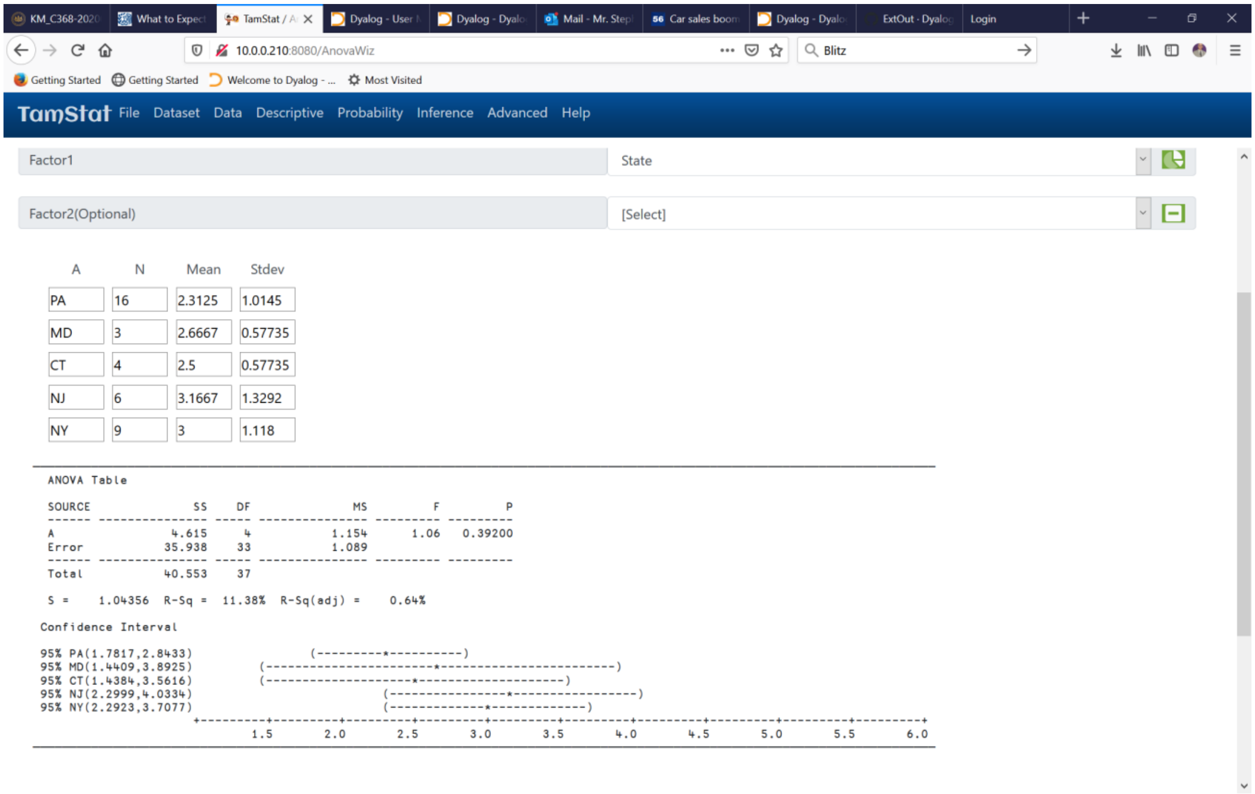
Operation: anova

Variable: HealthCare

Factor1: State

Factor2(Optional): [Select]

Scroll down to see the results which consist of summary statistics for each category, an Anova table and confidence intervals for each group mean: (See next page)



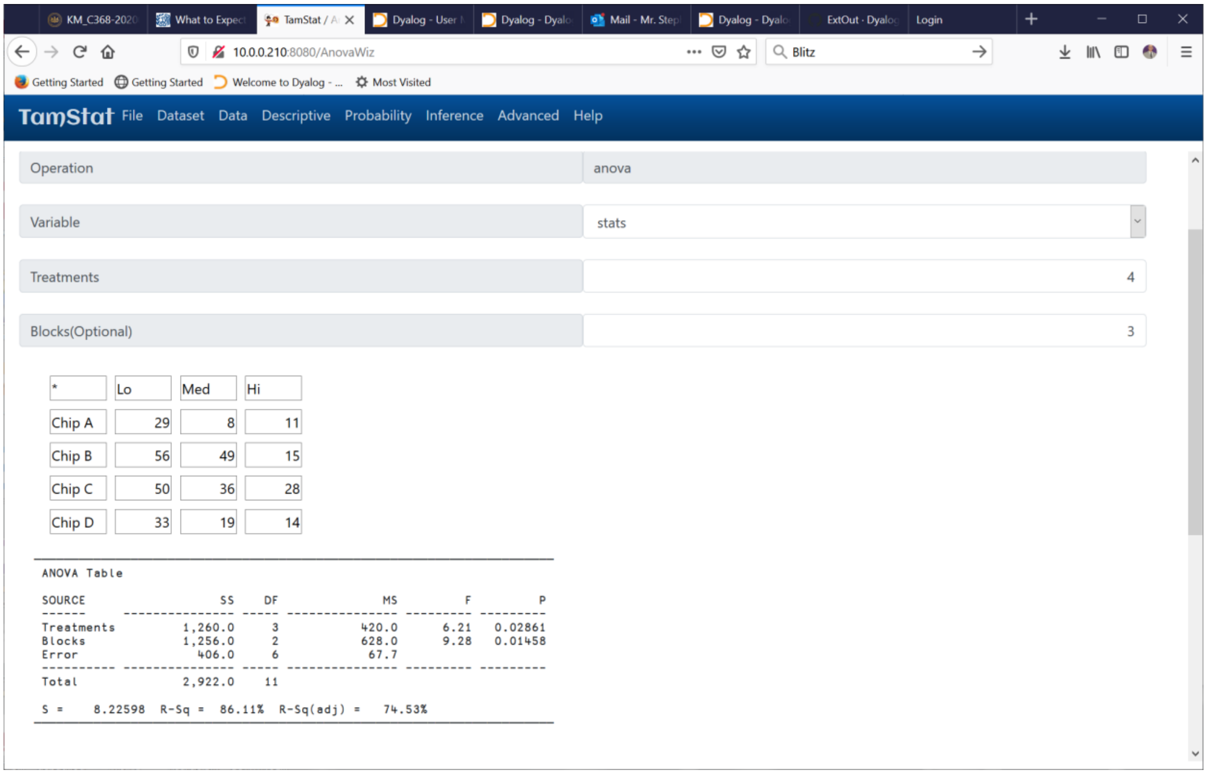
Factor1: State

Factor2(Optional): [Select]

A	N	Mean	Stdev
PA	16	2.3125	1.0145
MD	3	2.6667	0.57735
CT	4	2.5	0.57735
NJ	6	3.1667	1.3292
NY	9	3	1.118

```
 ANOVA Table

 SOURCE              SS    DF              MS         F         P
 ------ --------------- ----- --------------- --------- ---------
 A                4.615     4           1.154      1.06   0.39200
 Error           35.938    33           1.089
 ------ --------------- ----- --------------- --------- ---------
 Total           40.553    37

 S =    1.04356  R-Sq =  11.38%  R-Sq(adj) =    0.64%

Confidence Interval

95% PA(1.7817,2.8433)                (---------*----------)
95% MD(1.4409,3.8925)         (-----------------------*-------------------------)
95% CT(1.4384,3.5616)         (--------------------*---------------------)
95% NJ(2.2999,4.0334)                          (-----------------*-----------------)
95% NY(2.2923,3.7077)                          (--------------*-------------)
                     +---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
                             1.5       2.0       2.5       3.0       3.5       4.0       4.5       5.0       5.5       6.0
```

To handle blocked data, select “stats” for variable, select the number of treatments and blocks, and populate the table for each treatment combination:

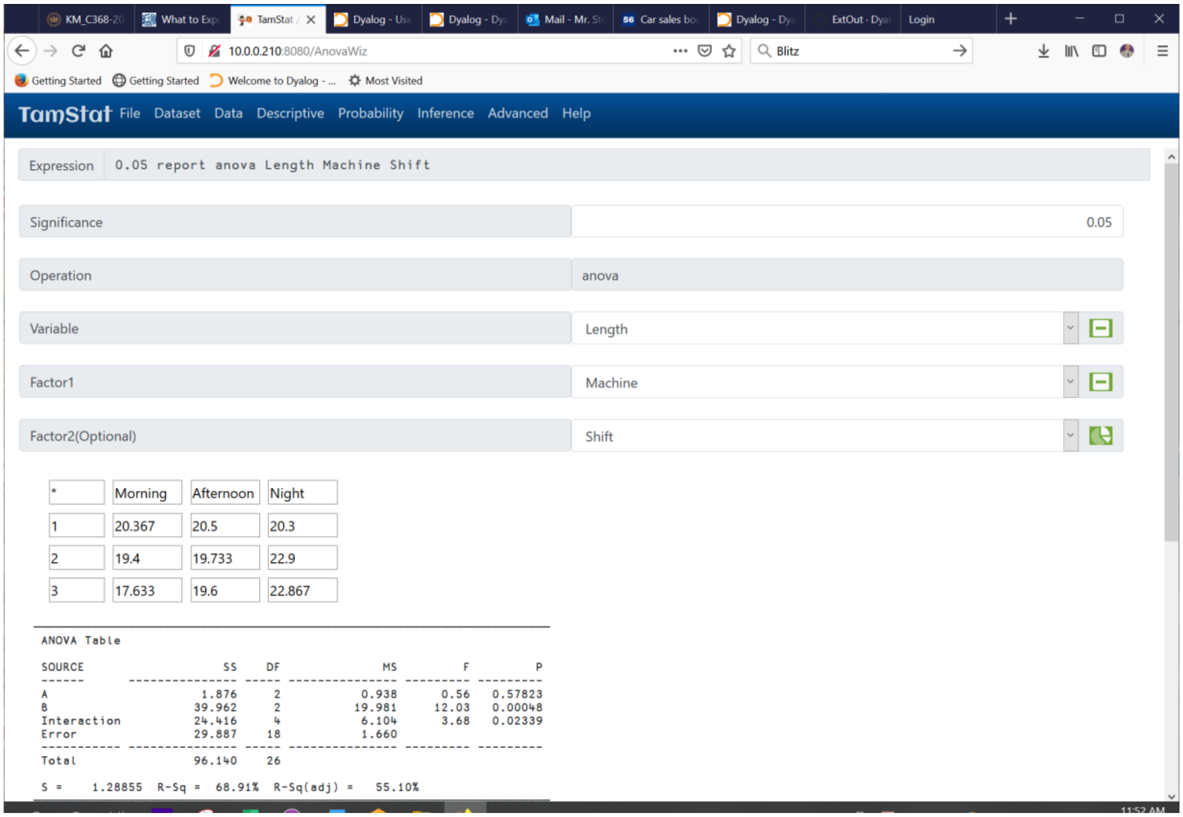

TamStat File Dataset Data Descriptive Probability Inference Advanced Help
10.0.0.210:8080/AnovaWiz

Operation	anova
Variable	stats
Treatments	4
Blocks(Optional)	3

*	Lo	Med	Hi
Chip A	29	8	11
Chip B	56	49	15
Chip C	50	36	28
Chip D	33	19	14

```
ANOVA Table

SOURCE                    SS     DF                MS          F          P
------        --------------- ----- ---------------- --------- ---------
Treatments           1,260.0      3             420.0       6.21   0.02861
Blocks               1,256.0      2             628.0       9.28   0.01458
Error                  406.0      6              67.7
---------- --------------- ----- ---------------- --------- ---------
Total                2,922.0     11

S =    8.22598  R-Sq =  86.11%  R-Sq(adj) =   74.53%
```

To perform a two-way ANOVA, you need a balanced design.   The PP database in Section 4.2.3 can be used: Select the variable “Length” , and select “Machine” for Factor 1 and “Shift” for factor 2.   The table will automatically be populated with means for each treatment combination. To ensure a balanced design, the number of replications for each treatment combination in the database must be the same.


TamStat File Dataset Data Descriptive Probability Inference Advanced Help
10.0.0.210:8080/AnovaWiz

Expression	0.05 report anova Length Machine Shift
Significance	0.05
Operation	anova
Variable	Length
Factor1	Machine
Factor2(Optional)	Shift

*	Morning	Afternoon	Night
1	20.367	20.5	20.3
2	19.4	19.733	22.9
3	17.633	19.6	22.867

```
ANOVA Table

SOURCE                    SS     DF                MS          F          P
------          -------------- ----- ---------------- --------- ---------
A                      1.876      2             0.938       0.56   0.57823
B                     39.962      2            19.981      12.03   0.00048
Interaction           24.416      4             6.104       3.68   0.02339
Error                 29.887     18             1.660
---------- -------------- ----- ---------------- --------- ---------
Total                 96.140     26

S =    1.28855  R-Sq =  68.91%  R-Sq(adj) =   55.10%
```

4.10 – Exercises

1. A market research company randomly divides 12 stores from a large grocery chain into three groups of four stores each in order to compare the effect on mean sales of three different types of displays. The company uses display type 1 in four of the stores, display type 2 in four others, and display type 3 in the remaining four stores. Then it records the amount of sales (in $1,000's) during a one month period at each of the twelve stores. The table shown below reports the sales information.

Obs	**Display Type I**	**Display Type II**	**Display Type III**
1	108	135	160
2	117	125	150
3	135	135	136
4	115	120	139
Total	475	515	585

Chapter 5 - Regression

Suppose you are having a wedding. It costs \$500 to rent the hall plus \$100 per guest. Using this information answer the following two questions:

1. If you have 30 guests, how much does it cost to hold the wedding?
2. You have a budget of \$4000. How many guests can you invite?

We can answer the first question using simple arithmetic:

```
      500+30×100
3500
```

Question #2 is a little more difficult. We first write the following equation:

$$y = 500 + 100x$$

Then we solve the equation for x using high-school algebra:

$$x = \frac{y - 500}{100}$$

Now, we can calculate the result by substituting $y = 4000$:

```
      (4000-500)÷100
35
```

Now try to answer the following question:

3. If the wedding hall advertises the following fee schedule, what are the fixed costs and variable costs?

Number of Guests	Total Cost
20	\$2,500
50	\$5,500

Before we attempt to solve question #3, let's backtrack a little and look at all three questions. We are dealing with a linear equation of the form: $y = b_0 + b_1 x$ where b_0 is the y-intercept or fixed cost, and b_1 is the slope or variable cost. Using functional notation, we can write this as: $y = f(x)$ to show that the total cost (y) is a function of the number of guests (x). In question #1, we know x and we apply the function $f(x) = 500 + 100x$ to find y. In other words $? = f(x)$ In question #2, we know y and we have to determine the inverse function to find x: $y = f(?)$ $x = f^{-1}(y)$. In question #3, we know both x and y; we are trying to find the function *f*! $y = ?(x)$ Since we know *f*(x)is a linear function, all we need to do is find the intercept b_0 and slope b_1 to completely determine *f*. Thus from Question #3, we can form two equations in two unknowns:

$$2500 = b_0 + 20b_1$$
$$5500 = b_0 + 50b_1$$

By subtracting the first equation from the second we can find the solution:

$$3000 = 30b_1$$

Thus $b_1 = 100$ and substituting this back into the first equation will allow us to find $b_0 = 500$.

This works well when you have two equations. But suppose you have three or more? Let's append our fee schedule:

Number of Guests	Total Cost
20	\$ 2,500
50	\$ 5,500
100	\$10,000

Now we have three equations in two unknowns. Most of the time we will not be able to find a line that goes through all three points. However, we can find the best fitting line using a technique known as linear regression.

5.1 Simple Regression

A car dealer runs television ads for five weeks and records the number of cars sold that week:

Week	Television Ads Run	Cars Sold
1	**1.0**	**14.0**
2	**3.0**	**24.0**
3	**2.0**	**18.0**
4	**1.0**	**17.0**
5	**3.0**	**27.0**

The number of ads run for 5 weeks is known as the independent or predictor variable *X*.

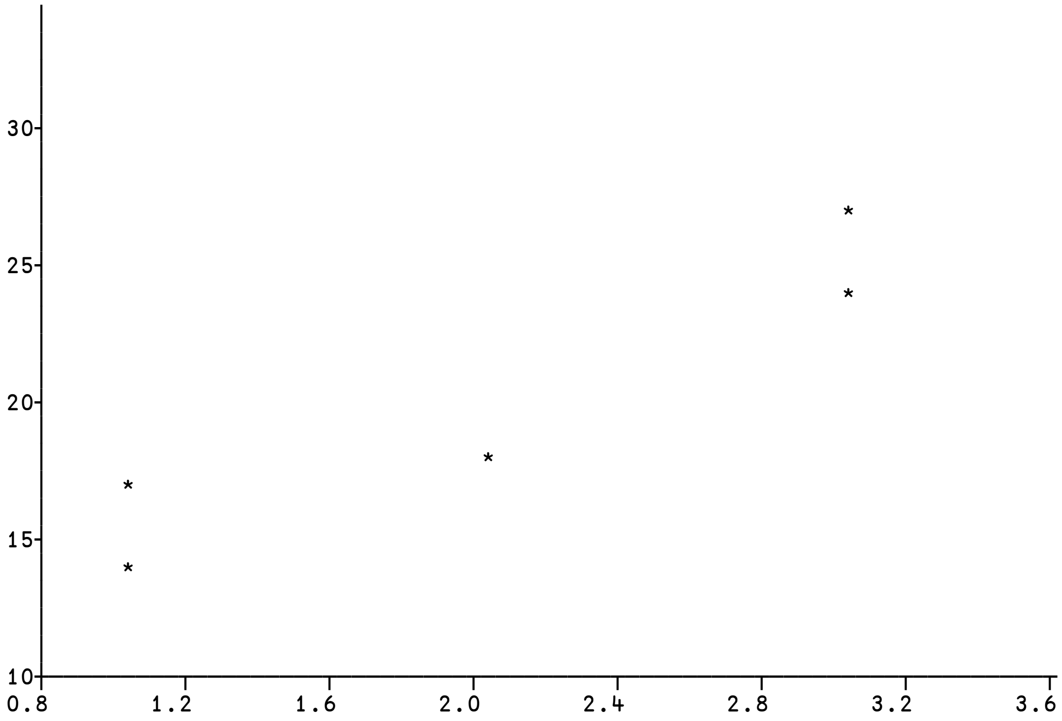
```
      ADS←1 3 2 1 3
```

The number of cars sold in each of the 5 weeks is known as the dependent or response variable *Y*.

```
      CARS←14 24 18 17 27
```

First, we generate a scatterplot. Note that the points are not exactly in a straight line:

```
      G←CARS scatterPlot ADS
      G.Output
```



The `regress` function produces a namespace `RG` containing various outputs:

```
      RG←CARS regress ADS
```

The estimates for the intercept b_0 and the slope b_1 are represented by the vector `B`.

```
      B0 B1←RG.B             ⍝ Intercept, Slope
10 5
```

The regression model is: $\hat{Y} = b_0 + b_1X$. When we enter a particular value for the predictor variable *X* we get an estimate for the response variable Y:

```
     B0+B1×2        ⍝ Estimated cars sold given 2 ads
20
```

This model predicts that you will sell about 20 cars if you run two ads. The intercept indicates you will probably sell 10 cars without advertising, and the slope tells you that the marginal benefit of each additional ad is 5 more cars sold.

The residual is the difference between the predicted value and the actual value. It tells us how far our prediction was away from the true value: $e_i = y_i - \hat{y}_i$

```
      18-20                          ⍝ Residual
¯2
      YHAT←B0+B1×ADS                 ⍝ All estimates
      CARS-YHAT                      ⍝ All residuals
¯1 ¯1 ¯2 2 2
```

The sum of the residuals is 0, since some estimates will be above the true value and others will be below.

```
      sum CARS-YHAT
0
```

We would like to find the size of the average error. So, we will square the differences and divide by the degrees of freedom, which for simple regression is $n - 2$.

```
      sumSquares CARS-YHAT
14
      sqrt 14 div 5-2
2.1602
```

Fortunately, we can obtain this result directly from the regression object `R` we created:

```
      RG.S
2.1602
```

To obtain the sample correlation, we can simply apply the correlation function.

```
     CARS corr ADS
0.93659
```

To obtain R-Squared we square this result:

```
      0.93659*2
0.8772
```

The regression object gives us this value directly:

```
      RG.RSq
0.87719
```

Since our sample size is small, we may want to perform a hypothesis test on the correlation coefficient:

```
      report CARS ADS corr hypothesis = 0
```

```
 r =0.93659
 n =5
 Standard Error: 0.20233

 Hypothesis Test

  H₀: ρ=0              H₁: ρ≠0
 ┌──────────────────┬───────────────────┐
 │Test Statistic:   │P-Value:           │
 │t=4.6291          │p=0.018986         │
 ├──────────────────┼───────────────────┤
 │Critical Value:   │Significance Level:│
 │t(α/2;df=3)=3.1824│α=0.05             │
 └──────────────────┴───────────────────┘
  Conclusion: Reject H₀
```

Finally, we can obtain a full report of the regression using the `report` function:

```
      report RG
```

```
 The regression equation is:

 Y←10+(5×X1)+E

  ANOVA Table

  SOURCE                 SS    DF              MS         F         P
  ------     --------------- ----- --------------- --------- ---------
  Regression         100.00     1          100.00     21.43   0.01899
  Error               14.00     3            4.67
  ---------- --------------- ----- --------------- --------- ---------
  Total              114.00     4

  S =    2.16025  R-Sq =  87.72%  R-Sq(adj) =   83.63%

 Solution

   Variable     Coeff        SE         T         P
 Intercept      10.00      2.37   4.22577   0.02424
 B1              5.00      1.08   4.62910   0.01899
```

Now let’s see how to use data directly from an Excel spreadsheet. We will use the variables from the class data we imported earlier.

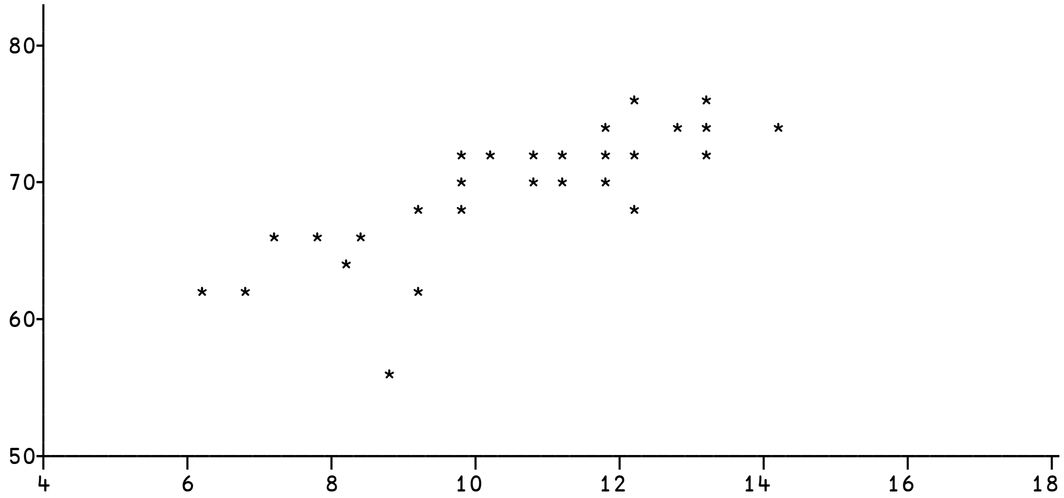
CSI Scranton: You are investigating a murder, and you find a bloody footprint near the victim. When you measure it, it matches a size 9-1/2 shoe. How tall is the suspect?

Since there are two variables involved: `Height` and `ShoeSize`, we see that there is a relationship between them by calculating their correlation:

```
      #.SD.Height corr #.SD.ShoeSize
0.82689
```

We can look at a scatter plot and can see that the relationship is positive, linear and moderately strong:

```
      #.SD.Height scatterPlot #.SD.ShoeSize
```



To find the height of the suspect, you need to establish a relationship between shoe size and height by finding a linear model. We again use the regress function and give it the name of our database object `D` followed by a listed of variable names delimited by spaces:

```
      MODEL←#.SD.Height  regress #.SD.ShoeSize
```

To find the regression coefficients, we simply enter

```
      MODEL.B
50.77060572 1.771435553
```

To find the estimated height we could enter:

```
      MODEL.(B[0] + B[1] × 9.5)
67.599
```

But it's much simpler to apply the generated linear function:

```
      MODEL.f 9.5
67.59924348
```

To obtain a full report enter the following:

```
     report MODEL
```

```
 The regression equation is:

 Height←50.771+(1.7714×ShoeSize)+E

  ANOVA Table

  SOURCE                  SS     DF              MS         F         P
  ------     --------------- ----- --------------- --------- ---------
  Regression          506.89      1          506.89     77.83  <0.00001
  Error               234.46     36            6.51
  ---------- --------------- ----- --------------- --------- ---------
  Total               741.35     37

  S =     2.55199  R-Sq =  68.37%  R-Sq(adj) =   67.50%

 Solution

 Variable       Coeff         SE          T         P
 Intercept      50.77       2.08   24.37953  <0.00001
 ShoeSize        1.77       0.20    8.82224  <0.00001
```

To find a 95% confidence interval for the height of the average person with a size 9 ½ shoe, enter:

```
     MODEL.f confInt 9.5
66.717 68.481
```

To find a 95% prediction interval for the height of the perpetrator we enter:

```
     MODEL.f predInt 9.5
62.349 72.85
```

We can say that the suspect is between 62 and 73 inches tall. Notice that this interval is much wider than the confidence interval. If we reduce the confidence level to 90%, we can narrow the interval.

```
     0.9 MODEL.f predInt 9.5
63.229 71.97
```

5.2 Multiple Regression

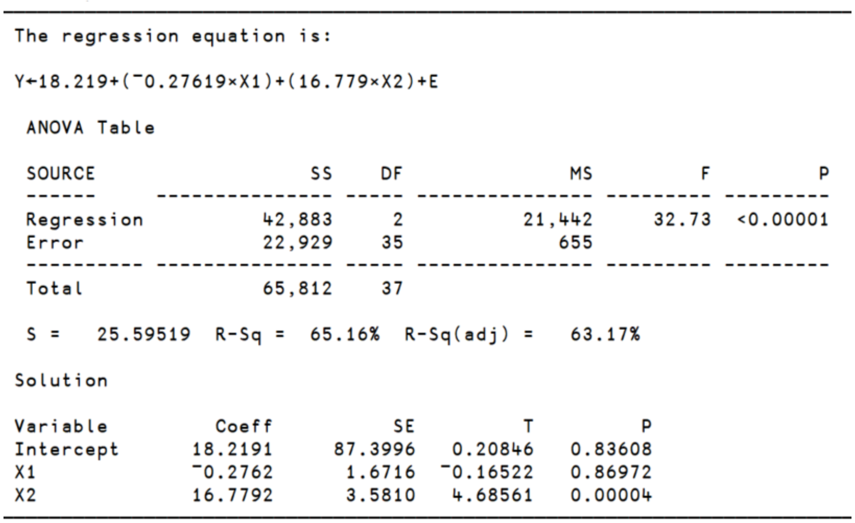
Sometimes it is useful to have more than one predictor variable. For example, we can estimate height from both weight and shoe size. The two-independent-variable model is:

$$\hat{Y} = b_0 + b_1X_1 + b_0X_2 \qquad \text{or} \qquad Y = b_0 + b_1X_1 + b_0X_2 + \varepsilon$$

where ε represents the residual or error.

Multiple regression in TamStat requires the right argument to take on one of three forms: a variable list, a matrix or a namespace. For both the variable list and matrix forms, the left argument the response variable, a numeric vector.

```
MODEL←Weight regress Height ShoeSize ⍝  Variable List
XX←Height,⍪ShoeSize                  ⍝  XX is N×2 Matrix
MODEL←Weight regress XX              ⍝  Multiple Regression
report MODEL
```

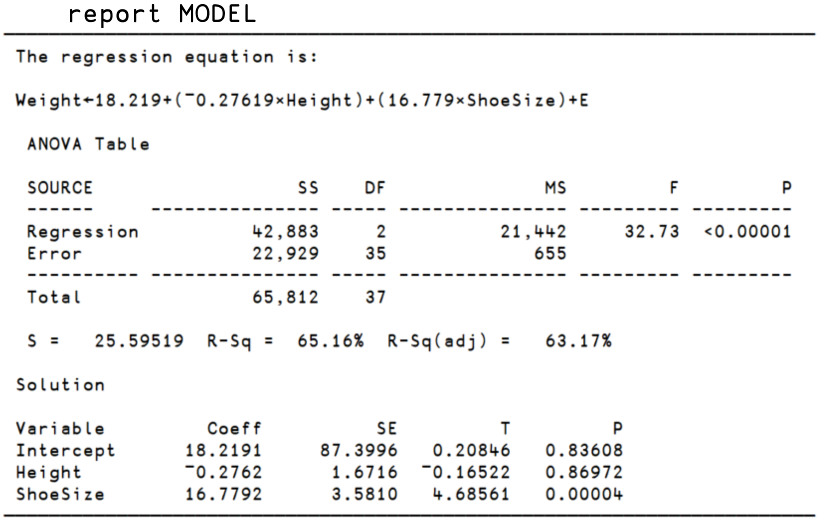


```
The regression equation is:

Y←18.219+(¯0.27619×X1)+(16.779×X2)+E

 ANOVA Table

 SOURCE                  SS    DF              MS         F         P
 ------     --------------- ----- --------------- --------- ---------
 Regression          42,883     2          21,442     32.73  <0.00001
 Error               22,929    35             655
 ---------- --------------- ----- --------------- --------- ---------
 Total               65,812    37

 S =   25.59519  R-Sq =  65.16%  R-Sq(adj) =   63.17%

Solution

Variable        Coeff          SE         T         P
Intercept     18.2191     87.3996   0.20846   0.83608
X1            ¯0.2762      1.6716  ¯0.16522   0.86972
X2            16.7792      3.5810   4.68561   0.00004
```

Observe that the intercept and the coefficient for shoe size are significant, the weight coefficient is not significant due to the large p-value. This shows that `Weight` does not contribute significantly to height when shoe size is in the model.

To preserve the names of the variables, one can use a namespace which represents a database:

```
V←'Weight Height ShoeSize'  ⍝ Variables of interest
DB←V selectFrom SD          ⍝ Put variables into namespace
MODEL←'Weight' regress DB   ⍝ Left argument is name of response variable.
report MODEL
```



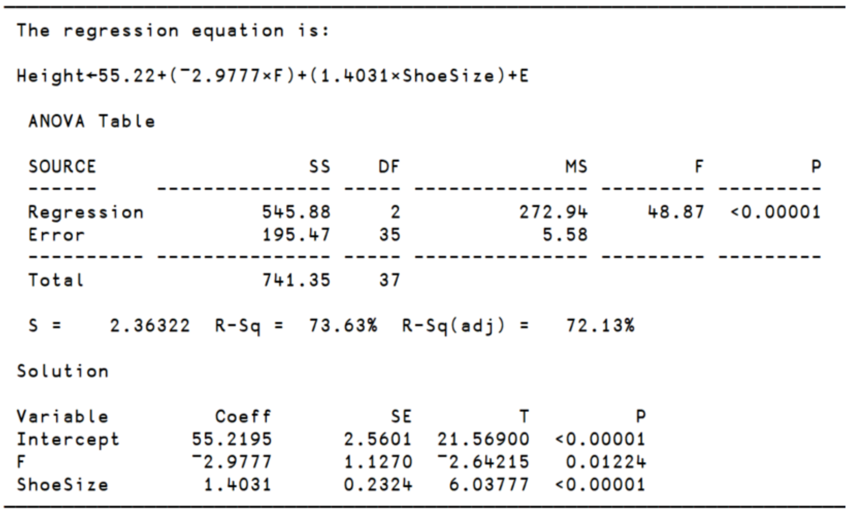
```
The regression equation is:

Weight←18.219+(¯0.27619×Height)+(16.779×ShoeSize)+E

 ANOVA Table

 SOURCE                  SS    DF              MS         F         P
 ------     --------------- ----- --------------- --------- ---------
 Regression          42,883     2          21,442     32.73  <0.00001
 Error               22,929    35             655
 ---------- --------------- ----- --------------- --------- ---------
 Total               65,812    37

 S =   25.59519  R-Sq =  65.16%  R-Sq(adj) =   63.17%

Solution

Variable        Coeff          SE         T         P
Intercept     18.2191     87.3996   0.20846   0.83608
Height        ¯0.2762      1.6716  ¯0.16522   0.86972
ShoeSize      16.7792      3.5810   4.68561   0.00004
```

```
      MODEL.f 68 9.5               ⍝ Estimate weight from height, shoe size
158.84
      0.9 MODEL.f confInt 68 9.5   ⍝ 90% Confidence interval
151.38 166.3
      0.9 MODEL.f predInt 68 9.5   ⍝ 90% Prediction interval
114.96 202.72
```

5.3 Indicator Variables

If there are character fields in a database, TamStat treats them as indicator variables. If there are more than two categories, TamStat will create multiple indicator variables. The indicator variable names will be taken from the value(s) of the indicator variable. There will always be $k - 1$ indicator variables when there are k unique values in the character field.

```
V←'Height ShoeSize Sex'     ⍝ “Sex” is a character field
DB←V selectFrom D           ⍝ Create a namespace
MODEL←'Height' regress DB   ⍝ “Height” is response variable
report MODEL                ⍝ “F” is a value from “Sex” character field
```

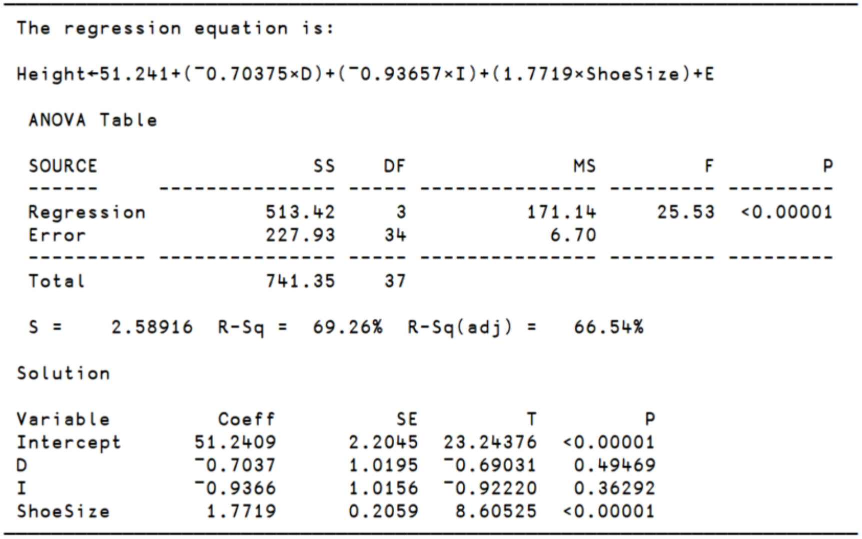


```
The regression equation is:

Height←55.22+(¯2.9777×F)+(1.4031×ShoeSize)+E

 ANOVA Table

 SOURCE                  SS    DF              MS         F         P
 ------     --------------- ----- --------------- --------- ---------
 Regression          545.88     2          272.94     48.87  <0.00001
 Error               195.47    35            5.58
 ---------- --------------- ----- --------------- --------- ---------
 Total               741.35    37

 S =    2.36322  R-Sq =  73.63%  R-Sq(adj) =   72.13%

Solution

Variable         Coeff          SE          T         P
Intercept      55.2195      2.5601   21.56900  <0.00001
F              ¯2.9777      1.1270   ¯2.64215   0.01224
ShoeSize        1.4031      0.2324    6.03777  <0.00001
```

5.3.1 Indicator Variables with more than two Categories

Let’s replace Sex with Party. There are 3 parties: Democrat, Independent and Republican. TamStat will create two indicator variables from Party: “Democrat” and “Independent”. “Republican” will be the base case.

```
V←'Height ShoeSize Party'   ⍝ Replace “Sex” with “Party”
DB←V selectFrom D           ⍝ Create new database
MODEL←'Height' regress DB   ⍝ Height is response variable
report MODEL                ⍝ “Republican” is base Case
```

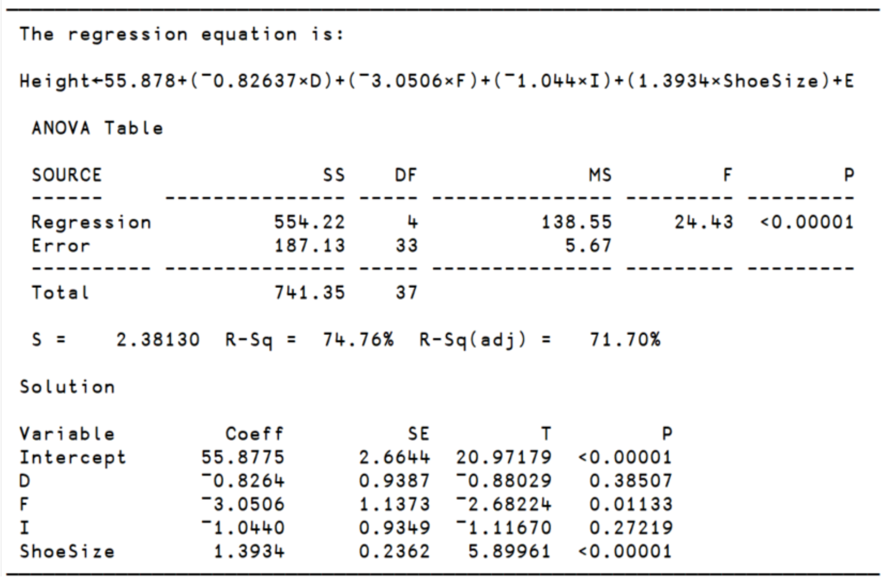


```
The regression equation is:

Height←51.241+(¯0.70375×D)+(¯0.93657×I)+(1.7719×ShoeSize)+E

 ANOVA Table

 SOURCE                  SS    DF              MS         F         P
 ------     --------------- ----- --------------- --------- ---------
 Regression          513.42     3          171.14     25.53  <0.00001
 Error               227.93    34            6.70
 ---------- --------------- ----- --------------- --------- ---------
 Total               741.35    37

 S =    2.58916  R-Sq =  69.26%  R-Sq(adj) =   66.54%

Solution

Variable         Coeff          SE          T         P
Intercept      51.2409      2.2045   23.24376  <0.00001
D              ¯0.7037      1.0195   ¯0.69031   0.49469
I              ¯0.9366      1.0156   ¯0.92220   0.36292
ShoeSize        1.7719      0.2059    8.60525  <0.00001
```

5.3.2 Multiple Indicator Variables

When there are more than one character field in a database, the number of indicator variables will become $\sum_{i=1}^{m} k_i - m$ where m is the number of character fields:

```
 DB←'Height ShoeSize Party Sex' selectFrom SD ⍝ Two-character fields
 Report 'Height' regress DB                   ⍝ 3 Indicator Variables
```

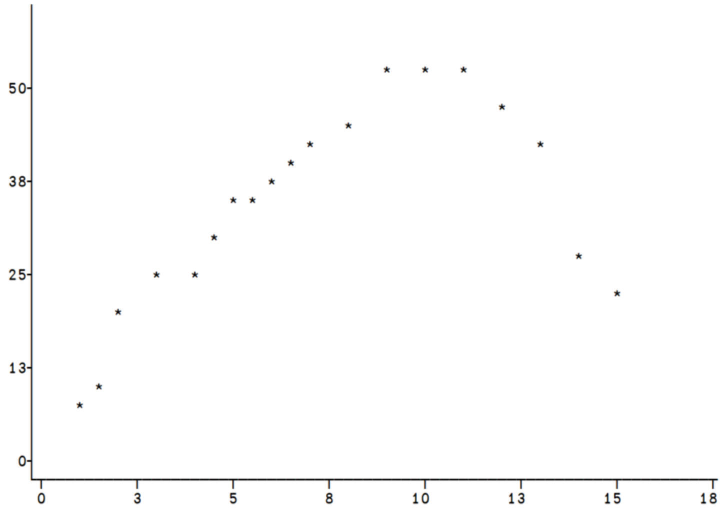


```
The regression equation is:

Height←55.878+(¯0.82637×D)+(¯3.0506×F)+(¯1.044×I)+(1.3934×ShoeSize)+E

 ANOVA Table

 SOURCE                   SS    DF              MS         F         P
 ------     --------------- ----- --------------- --------- ---------
 Regression           554.22     4          138.55     24.43  <0.00001
 Error                187.13    33            5.67
 ---------- --------------- ----- --------------- --------- ---------
 Total                741.35    37

 S =    2.38130  R-Sq =  74.76%  R-Sq(adj) =   71.70%

Solution

Variable        Coeff         SE         T         P
Intercept     55.8775     2.6644  20.97179  <0.00001
D             ¯0.8264     0.9387  ¯0.88029   0.38507
F             ¯3.0506     1.1373  ¯2.68224   0.01133
I             ¯1.0440     0.9349  ¯1.11670   0.27219
ShoeSize       1.3934     0.2362   5.89961  <0.00001
```

5.4 Polynomial Regression

We are trying to predict Y = tensile strength in p.s.i. from X = hardwood concentration (%) .

```
    X←1 1.5 2 3 4 4.5 5 5.5 6 6.5 7 8 9 10 11 12 13 14 15
    Y←6.3 11.1 20 24 26.1 30 33.8 34 38.1 39.9 42 46.1 53.1 52 52.5 48
     Y,← 42.8 27.8 21.9
     scatterplot show Y X ⍝ Data are non-linear
```

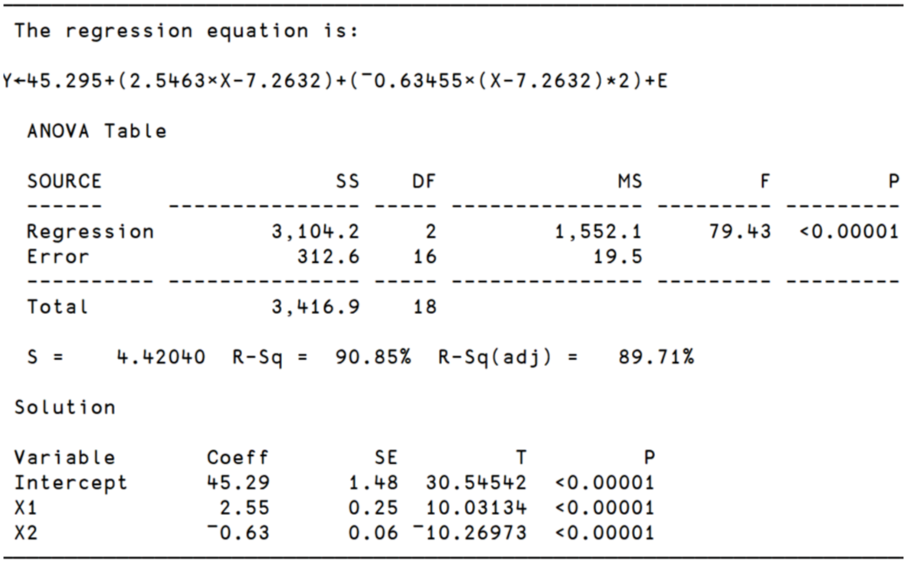


5.4.1 Quadratic Regression

Since the data above are clearly non-linear, perhaps a quadratic regression model would be appropriate. We use the base-value (⊥) or `poly` operand to indicate a quadratic model. Higher order polynomials are indicated by (n⊥) where n is degree of the polynomial. To eliminate correlation between the linear and squared terms, we correct by subtracting the mean:

$$y = b_0 + b_1(x - \bar{x}) + b_2(x - \bar{x})^2 + \varepsilon$$

```
   MODEL←Y poly regress X ⍝ Quadratic regression
   MODEL.B                ⍝ Constant, linear and square coefficients
45.295 2.5463 ¯0.63455
⍝  ↑       ↑       ↑
⍝ Int   Linear  Square
   MODEL.g 7 10 15        ⍝ The function g is the non-linear model:
44.581 47.511 27.012
   MODEL.g confInt 10     ⍝ Confidence Interval
44.402 50.619
   MODEL.g predInt 15
15.752 38.273
   report MODEL
```



```
 The regression equation is:

Y←45.295+(2.5463×X-7.2632)+(¯0.63455×(X-7.2632)*2)+E

  ANOVA Table

  SOURCE                  SS    DF              MS         F         P
  ------     --------------- ----- --------------- --------- ---------
  Regression         3,104.2     2         1,552.1     79.43  <0.00001
  Error                312.6    16            19.5
  ---------- --------------- ----- --------------- --------- ---------
  Total              3,416.9    18

  S =    4.42040  R-Sq =  90.85%  R-Sq(adj) =   89.71%

 Solution

 Variable       Coeff        SE         T         P
 Intercept      45.29      1.48  30.54542  <0.00001
 X1              2.55      0.25  10.03134  <0.00001
 X2             ¯0.63      0.06 ¯10.26973  <0.00001
```

5.4.2 Polynomial Models in Two or More Variables

We are trying to predict percentage yield from reaction time and temperature. Both predictor variables are quadratic, so we have a linear and quadratic term for each, plus an interactive term for a total of five input variables:

```
   PY←import 'processYield.CSV'
   X←PY.ReactionTime   ⍝ First Predictor Variable
76 80.5 78 89 93 92.1 77.8 84 87.3 75 85 90 85 79.2 83 82 94 91.4 95 81.1
88.8 91 87 86
   Y←PY.Temperature    ⍝ Second Predictor Variable
170 165 182 185 180 172 170 180 165 172 185 176 178 174 168 179 181 184 173
169 183 178 175 175
   Z←PY.Yield          ⍝ Response Variable
50.95 47.35 50.99 44.96 41.89 41.44 51.79 50.78 42.48 49.8 48.74 46.2 50.49
52.78 49.71 52.75 39.41 43.63 38.19 50.92 46.55 44.28 48.72 49.13
```

```
   MODEL←Z poly regress X Y
   MODEL.B    ⍝ Intercept, Linear, Quadratic and Interaction Coefficients
50.4 ¯0.72 ¯0.06 0.013 0.105 ¯0.038
⍝↑      ↑     ↑     ↑    ↑       ↑
⍝Int    X    X*2   X×Y   Y      Y*2

   MODEL.g 90 176         ⍝ Reaction time = 90 sec, temp = 176 degrees C
45.96
   MODEL.g confInt 90 176
45.481 46.439
   MODEL.g predInt 90 176
44.542 47.378
   report MODEL

The regression equation is:

Y←50.417+(¯0.71981×X1-85.467)+(¯0.059653×(X1-85.467)*2)+
  (0.012577×(X1-85.467)×(X2-175.79))+(0.10528×X2-175.79)+(¯0.037676×(X2-175.79)*2)+E

  ANOVA Table

  SOURCE                   SS    DF              MS         F         P
  ------     --------------- ----- --------------- --------- ---------
  Regression          416.31     5           83.26    206.28  <0.00001
  Error                 7.27    18            0.40
  ---------- --------------- ----- --------------- --------- ---------
  Total               423.58    23

  S =    0.63532  R-Sq =  98.28%  R-Sq(adj) =   97.81%

 Solution

 Variable       Coeff        SE         T         P
 Intercept      50.42      0.26 192.84947  <0.00001
 X1             ¯0.72      0.02 ¯29.36231  <0.00001
 X2             ¯0.06      0.00 ¯13.09424  <0.00001
 XY              0.01      0.01   2.40391   0.02721
 Y1              0.11      0.02   4.42554   0.00033
 Y2             ¯0.04      0.00  ¯9.18643  <0.00001
```

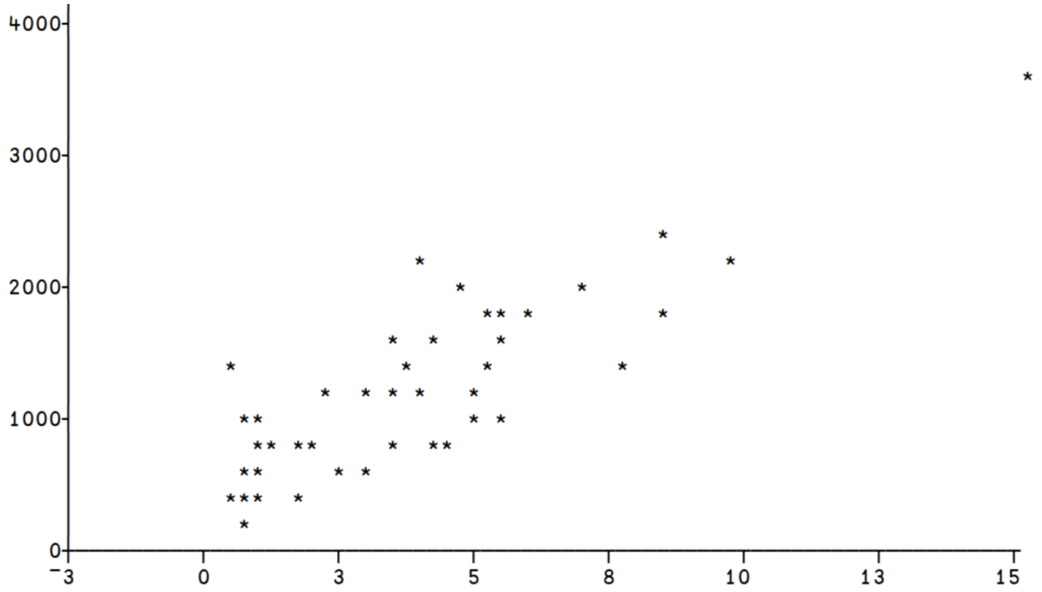
5.5 Variance Stabilizing Transformations

Linear regression assumes that the variance is constant regardless of the size of the response variable. In some cases this is not true. In order to compensate for this, we can transform the response variable to make the variance constant. There are several ways to do this. These are:

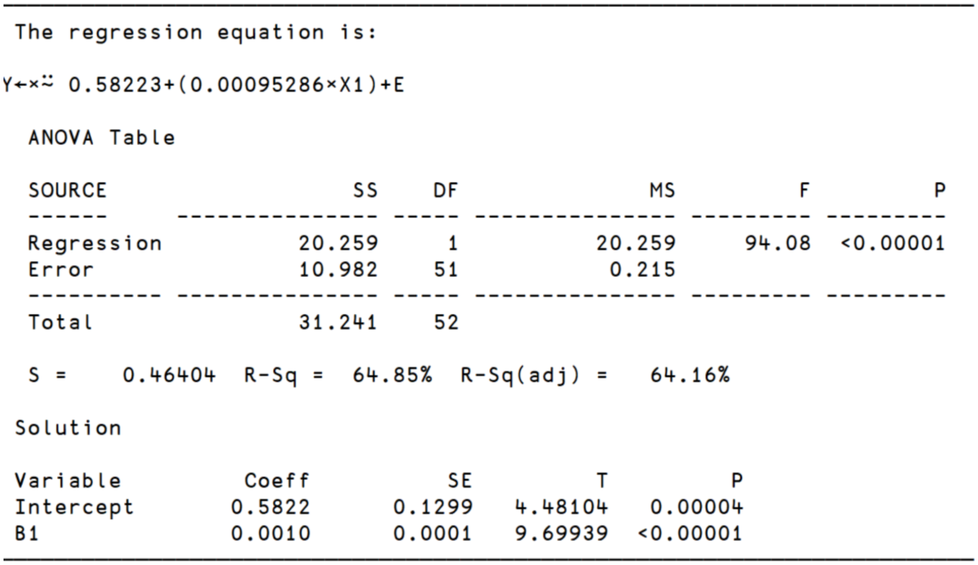
Variance proportional to	**Transformation**	**Left Operand to `regress`**
Constant	$y' = y$	`⊣`
$E(y)$	$y' = \sqrt{y}$	`sqrt`
$E(y)[1 - E(y)\ ]$	$y' = \sin^{-1}\sqrt{y}$	`arcsin∘sqrt`
$[E(y)]^2$	$y' = \ln y$	`ln`
$[E(y)]^3$	$y' = y^{-1/2}$	`÷∘sqrt`
$[E(y)]^4$	$y' = 1/y$	`÷`

An electric utility is developing a model relating peak demand to total monthly energy consumption. Data for 53 residential customers was collected. The following scatterplot shows that demand is proportional to its own variance. This indicates that we should use a square-root transform on the response variable.

```
    E←import 'Energy.csv'
    scatterPlot show E.(Usage Demand)
```



```
    E.Demand
0.79 0.44 0.56 0.79 2.7 3.64 4.73 9.5 5.34 6.85 5.84 5.21 3.25 4.43 3.16 0.5
0.17 1.88 0.77 1.39 0.56 1.56 5.28 0.64 4 0.31 4.2 4.88 3.48 7.58 2.63 4.99
0.59 8.19 4.79 0.51 1.74 4.1 3.94 0.96 3.29 0.44 3.24 2.14 5.71 0.64 1.9 0.51
8.33 14.94 5.11 3.85 3.93
    E.Usage
679 292 1012 493 582 1156 997 2189 1097 2078 1818 1700 747 2030 1643 414 354
1276 745 435 540 874 1543 1029 710 1434 837 1748 1381 1428 1255 1777 370 2316
1130 463 770 724 808 790 783 406 1242 658 1746 468 1114 413 1787 3560 1495
2221 1526
    MODEL←E.Demand sqrt regress E.Usage ⍝ Use Square Root transform
    MODEL.B                ⍝ Intercept and Slope
¯1.8313 0.0036828
    MODEL.f 1800           ⍝ Least Squares Estimate of transformed demand.
2.2974
    MODEL.g 1800           ⍝ Estimate demand in KW from  1800 KWH usage
5.2779
    MODEL.g confInt 1800   ⍝ Average demand range
4.4799 6.1413
    MODEL.g predInt 1800   ⍝ Individual demand range
1.8181 10.539

    report MODEL
```



```
 The regression equation is:

Y←×⍨ 0.58223+(0.00095286×X1)+E

  ANOVA Table

  SOURCE                  SS    DF              MS         F         P
  ------     --------------- ----- --------------- --------- ---------
  Regression          20.259     1          20.259     94.08  <0.00001
  Error               10.982    51           0.215
  ---------- --------------- ----- --------------- --------- ---------
  Total               31.241    52

  S =    0.46404  R-Sq =  64.85%  R-Sq(adj) =   64.16%

 Solution

 Variable         Coeff          SE         T         P
 Intercept       0.5822      0.1299   4.48104   0.00004
 B1              0.0010      0.0001   9.69939  <0.00001
```

In some cases, we may want to transform the predictor variable instead of the response variable. In that case, we preprocess the right argument before applying the regress operator. To generate energy from a windmill, we measure DC output and wind velocity. The physics of wind energy suggests that DC output approaches an upper limit of 2.5 which suggests the use of the following model: $y = \beta_0 + \beta_1/x + \varepsilon$.

```
  WV←import 'WindVelocity.csv'
  MODEL←WV.DCOutput regress ÷ WV.WindVelocity
  MODEL.B                 ⍝ Intercept, Slope of 1/x
2.9789 ¯6.9345
  MODEL.f ÷4 8 10         ⍝ Model approaches 2.5
1.2452 2.112 2.2854
  MODEL.f confInt÷10      ⍝ 95% Confidence interval of mean DC output at 10 mph
2.2284 2.3424
  MODEL.f predInt÷4 8 10 ⍝ 95% Prediction intervals of DC output
1.0453 1.4452
1.911  2.3131
2.0824 2.4884
```

5.6 Multiplicative Regression

Multiplicative regression models the following relationship: $y = ax^b$. This transforms both the predictor and response variables since it can be rewritten as: $\ln y = \ln a + b \ln x$. The ln operand to regress will only transform the response variable, so we use × as the operand to indicate multiplicative regression. The parameters will be the coefficient a and the exponent b.

As an example, let X and Y be the predictor and response variables:

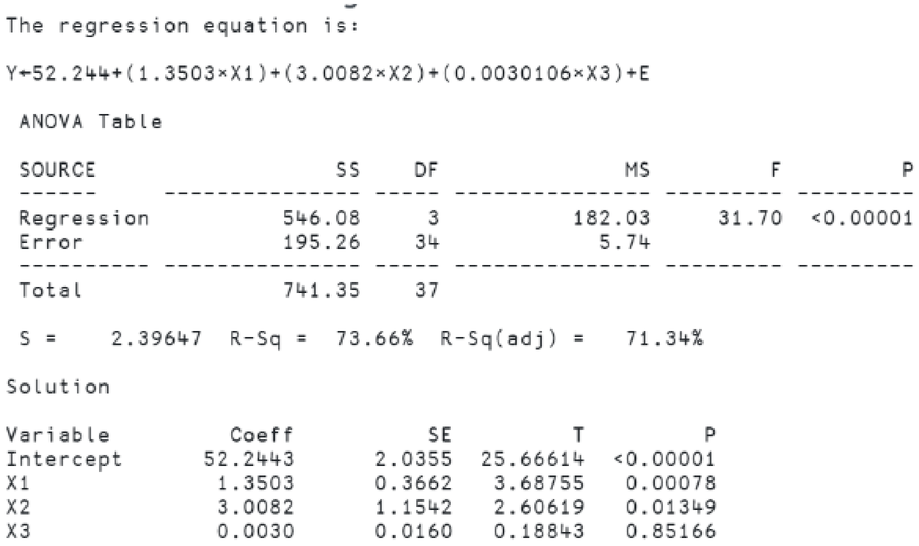
```
   X←14 14 8 10 6 7 5 10 5 13
   Y←864 870 83 176 37 50 8 164 26 584
   MODEL←Y ×regress X
   MODEL.B          ⍝ Coefficient A and Exponent B
0.029161 3.8539
   MODEL.g 9        ⍝ Estimate Y for X = 9
138.8
```

```
   Report MODEL

 The regression equation is:

Y←0.029161×(X1*3.8539)+E

  ANOVA Table

  SOURCE                    SS    DF              MS         F         P
  ------     --------------- ----- --------------- --------- ---------
  Regression          21.633     1          21.633    197.04  <0.00001
  Error                0.878     8           0.110
  ---------- --------------- ----- --------------- --------- ---------
  Total               22.511     9

  S =    0.33134  R-Sq =  96.10%  R-Sq(adj) =   95.61%

 Solution

 Variable          Coeff          SE         T         P
 Intercept       ¯3.5349      0.5991  ¯5.90046   0.00036
 B1               3.8539      0.2746  14.03708  <0.00001
```

5.6 Custom Designed Regression

The user may create a function which selects and/or transforms any of the variables in a database. This is particularly useful if there are multiple transformations. In order to do this one must create a transform function. The variable named Y becomes the response variable; `Int` defaults to 1; all others are predictor variables:

```
   makeTransFn 'Y←Height' 'X1←ShoeSize' 'X2←Sex eq ''M''' 'X3←Weight'
   MODEL←transform regress #.SD
   report MODEL
```



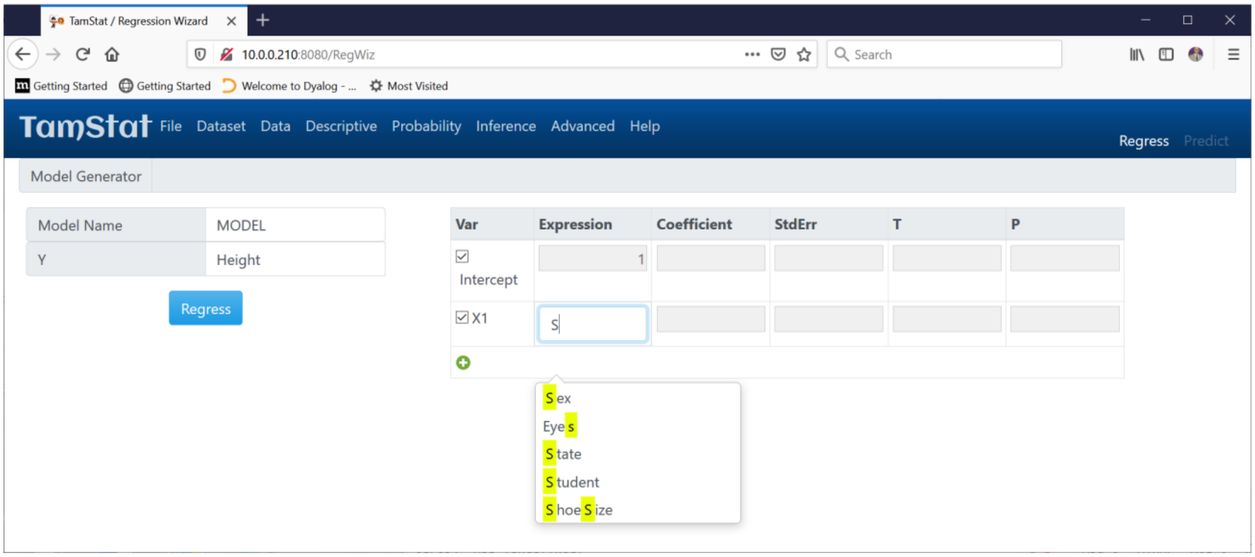
```
The regression equation is:

Y←52.244+(1.3503×X1)+(3.0082×X2)+(0.0030106×X3)+E

 ANOVA Table

 SOURCE                    SS    DF              MS         F         P
 ------     --------------- ----- --------------- --------- ---------
 Regression          546.08     3          182.03     31.70  <0.00001
 Error               195.26    34            5.74
 ---------- --------------- ----- --------------- --------- ---------
 Total               741.35    37

 S =    2.39647  R-Sq =  73.66%  R-Sq(adj) =   71.34%

Solution

Variable          Coeff          SE         T         P
Intercept       52.2443      2.0355  25.66614  <0.00001
X1               1.3503      0.3662   3.68755   0.00078
X2               3.0082      1.1542   2.60619   0.01349
X3               0.0030      0.0160   0.18843   0.85166
```

5.7 Regression Summary

<table>
<tr><th>Design</th><th>Left Argument</th><th>Operand / Function</th><th>Right Argument</th><th>Result</th></tr>
<tr><td>Simple linear regression</td><td>Response Variable*</td><td><code>regress</code></td><td>Predictor Variable</td><td>Intercept, Slope</td></tr>
<tr><td rowspan="3">Multiple Linear Regression</td><td rowspan="2">Response Variable*</td><td rowspan="2"><code>regress</code></td><td>Vector of Predictor Variables</td><td rowspan="3">Intercept, Coefficients for each predictor variable.</td></tr>
<tr><td>Matrix whose columns are predictor variables</td></tr>
<tr><td>Name of Response Variable* (Character string)</td><td><code>regress</code></td><td>Namespace containing all variables</td></tr>
<tr><td>Simple Quadratic Regression</td><td>Response Variable</td><td><code>⊥ regress</code></td><td>Predictor Variable</td><td>Intercept, Coeffiicents for centered data and squared centered data</td></tr>
<tr><td>Multiple Quadratic Regression</td><td>Response Variable</td><td><code>⊥ regress</code></td><td>Vector of Predictor Variables</td><td>Intercept, Linear, Quadratic and Interaction Coefficients</td></tr>
<tr><td>Polynomial Models</td><td>Response Variable</td><td><code>N ⊥ regress</code></td><td>Predictor Variable</td><td>Intercept, Coefficients for all powers up to N of predictor variable</td></tr>
<tr><td>Model with Indicator Variable(s)</td><td>Response Variable*</td><td><code>regress</code></td><td>Vector containing Predictor Variables and at least one Character Variable</td><td>Intercept, Coefficients for each predictor variable and $k-1$ coefficients for each character variable. (k = unique character values)</td></tr>
<tr><td>Variance Stabilizing Transformations</td><td>Response Variable</td><td><code><i>fn</i> regress</code>
<code>[ln|sqrt|÷|arcsin]</code></td><td>Predictor Variable</td><td>Intercept, Coefficients</td></tr>
<tr><td>Multiplicative Regression
$y = bx^a$</td><td>Response Variable</td><td><code>× regress</code></td><td>Predictor Variable</td><td>Constant, Powers</td></tr>
<tr><td>Indicator response variable</td><td>Boolean Variable</td><td><code>≠ regress</code></td><td>Predictor Variable</td><td>Intercept, Coefficients</td></tr>
<tr><td>Custom Regression</td><td>[None]</td><td><code>transform regress</code></td><td>Database (Namespace)</td><td>Intercept, Coefficients</td></tr>
<tr><td colspan="5">* Pseudo Left Argument – Actually, an array left operand.</td></tr>
</table>

5.8 Regression Wizard

To use the regression wizard, you must create predictor and response variables in the workspace or import them from a .csv file. Once you have done this, select "Regression" from the main menu. The following screen will appear. Indicate the Response Variable (Y) and the predictor Variable (X) by selecting from the dropdowns in each category:

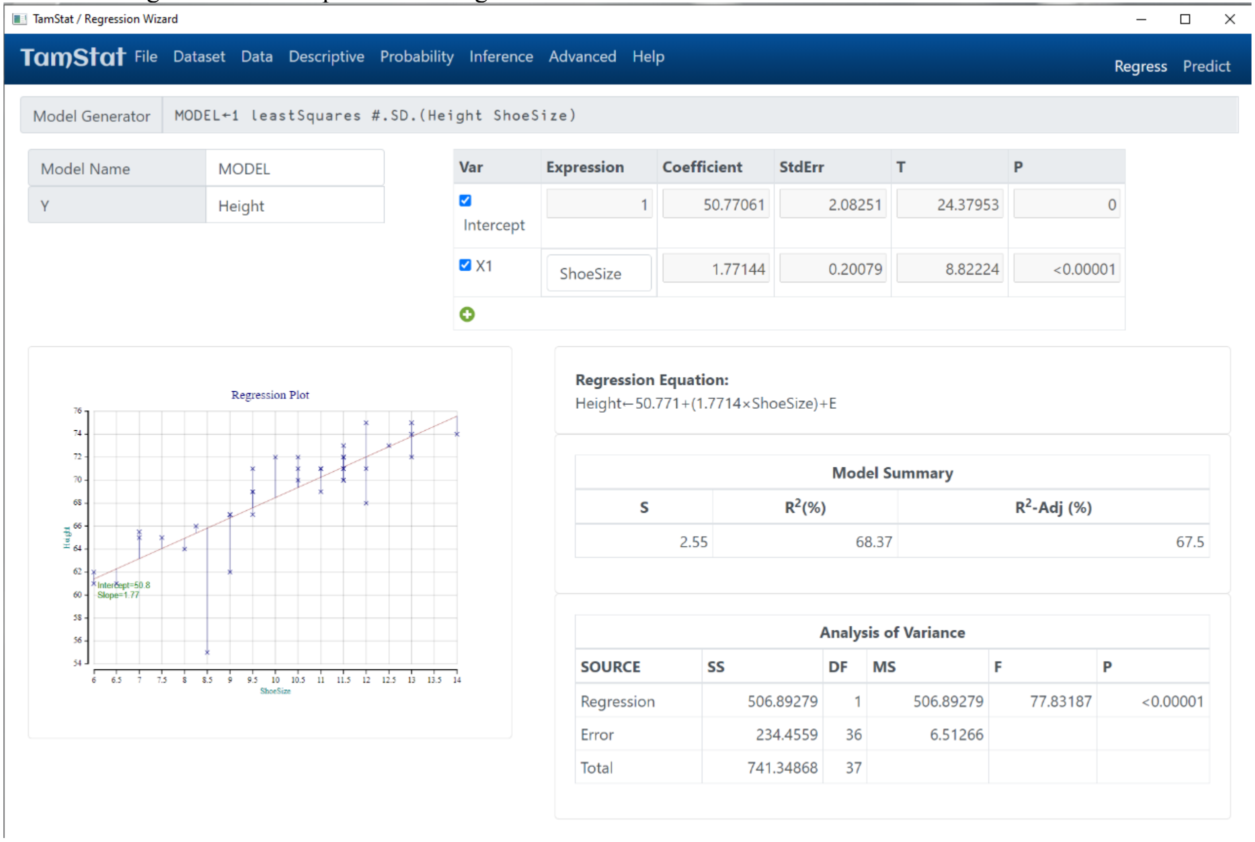


Model Name	MODEL
Y	Height

Var	Expression	Coefficient	StdErr	T	P
☑ Intercept	1				
☑ X1	S				

Sex
Eyes
State
Student
ShoeSize

Press the "Regress" button to perform the regression:



Model Generator `MODEL←1 leastSquares #.SD.(Height ShoeSize)`

Model Name	MODEL
Y	Height

Var	Expression	Coefficient	StdErr	T	P
☑ Intercept	1	50.77061	2.08251	24.37953	0
☑ X1	ShoeSize	1.77144	0.20079	8.82224	<0.00001

Regression Equation:
Height←50.771+(1.7714×ShoeSize)+E

Model Summary

S	R^2(%)	R^2-Adj (%)
2.55	68.37	67.5

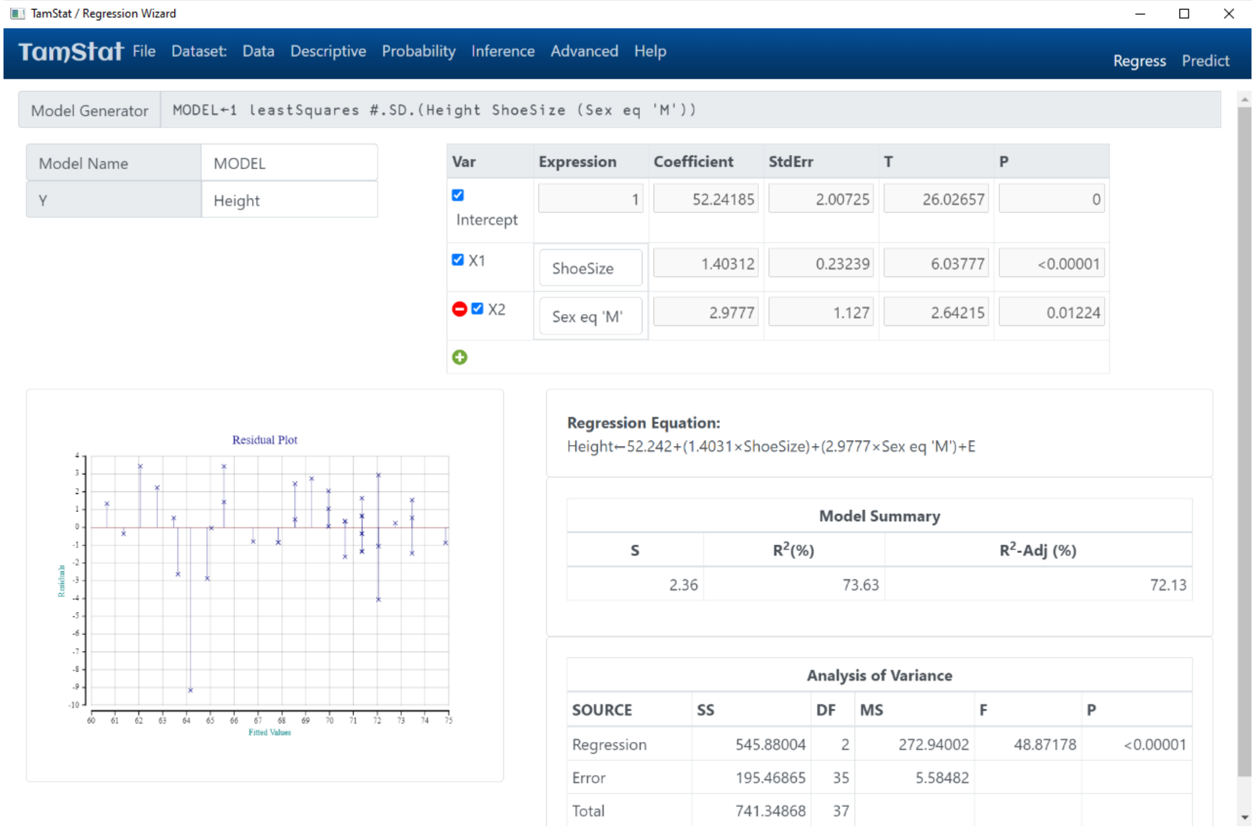
Analysis of Variance

SOURCE	SS	DF	MS	F	P
Regression	506.89279	1	506.89279	77.83187	<0.00001
Error	234.4559	36	6.51266		
Total	741.34868	37			

Note that there are six major areas on the regression screen.

1. The model generator: contains the TamStat expression.
2. The variable identification section. Here users enter the names of the response variable, and each predictor variable. In addition, there are checkboxes which allow the user to dynamically include or exclude predictor variables while building the model. The significance of the coefficients is also displayed here.
3. The regression plot: This will be a scatter plot for simple regression or a residual plot for multiple regression.
4. The regression equation
5. The model summary:
6. The ANOVA section which displays the analysis of variance and the overall F-Statistic

To perform multiple regression, select the plus sign below the variable "X1" to include additional variables in the model. Select the response variable and the predictor variables from the drop-down lists then select Regress from the menu. Note that you can use either a variable name or an expression for each of the variables:



Model Generator: `MODEL←1 leastSquares #.SD.(Height ShoeSize (Sex eq 'M'))`

Model Name	MODEL
Y	Height

Var	Expression	Coefficient	StdErr	T	P
☑ Intercept	1	52.24185	2.00725	26.02657	0
☑ X1	ShoeSize	1.40312	0.23239	6.03777	<0.00001
☑ X2	Sex eq 'M'	2.9777	1.127	2.64215	0.01224

Regression Equation:
Height←52.242+(1.4031×ShoeSize)+(2.9777×Sex eq 'M')+E

Model Summary

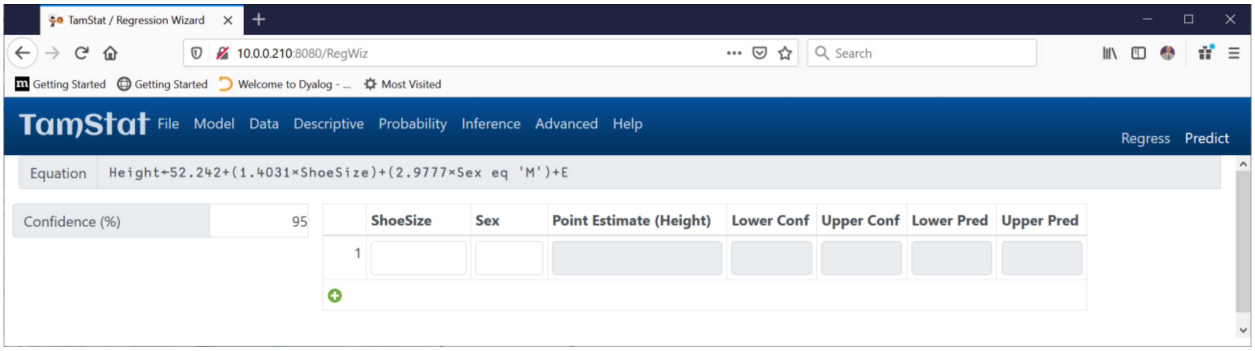
S	R^2(%)	R^2-Adj (%)
2.36	73.63	72.13

Analysis of Variance

SOURCE	SS	DF	MS	F	P
Regression	545.88004	2	272.94002	48.87178	<0.00001
Error	195.46865	35	5.58482		
Total	741.34868	37			

The coefficients are calculated along with the test statistic and p-value for each. Note the regression equation at the bottom of the screen which displayed as an APL expression, but which can also be interpreted as a mathematical expression. The APL expression to generate this model is also displayed.

To use the model to estimate the value of the dependent variable, select predict from the menu. The following screen will appear:
To use the model to predict Height, select "Predict" on the far right of the screen:

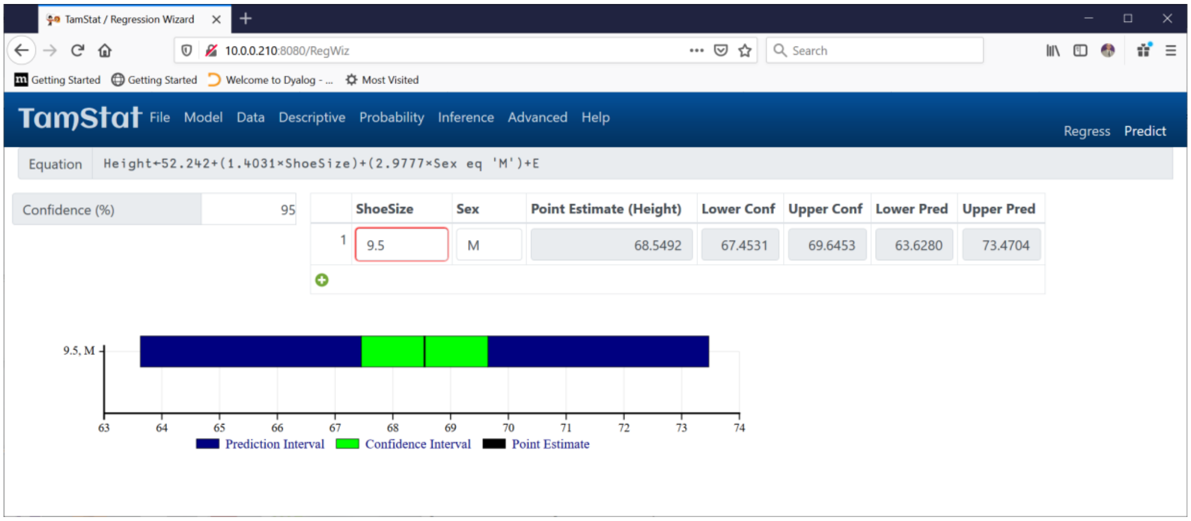


Equation `Height←52.242+(1.4031×ShoeSize)+(2.9777×Sex eq 'M')+E`

Confidence (%) 95

	ShoeSize	Sex	Point Estimate (Height)	Lower Conf	Upper Conf	Lower Pred	Upper Pred
1							

Enter “ShoeSize” and “Sex” of the “perpetrator” to get an estimate for the height of the suspect:

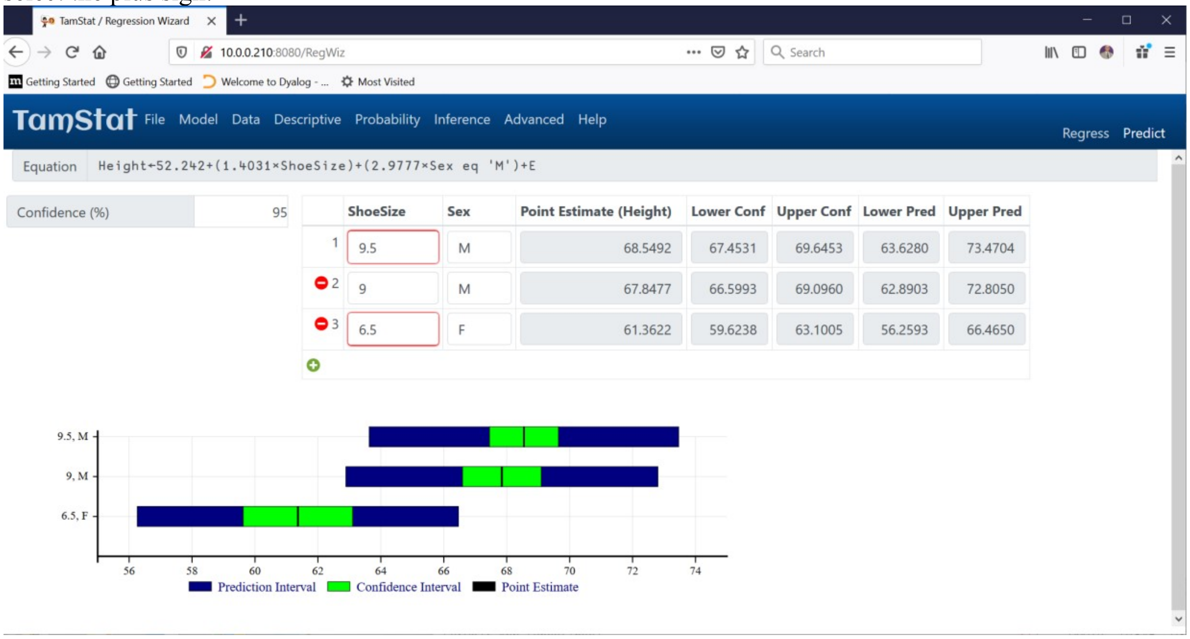


Equation `Height←52.242+(1.4031×ShoeSize)+(2.9777×Sex eq 'M')+E`

Confidence (%) 95

	ShoeSize	Sex	Point Estimate (Height)	Lower Conf	Upper Conf	Lower Pred	Upper Pred
1	9.5	M	68.5492	67.4531	69.6453	63.6280	73.4704

Note that the point estimate, 68.54 inches, also appears in the upper right-hand corner in the last column of the Y variable. Note the 95% prediction interval is between 63.63 and 73.47 inches. To do multiple predictions, simply select the plus sign:



Equation `Height←52.242+(1.4031×ShoeSize)+(2.9777×Sex eq 'M')+E`

Confidence (%) 95

	ShoeSize	Sex	Point Estimate (Height)	Lower Conf	Upper Conf	Lower Pred	Upper Pred
1	9.5	M	68.5492	67.4531	69.6453	63.6280	73.4704
2	9	M	67.8477	66.5993	69.0960	62.8903	72.8050
3	6.5	F	61.3622	59.6238	63.1005	56.2593	66.4650

5.9 Exercises

1. The following table lists the number of years of use and the prices (in thousands of dollars) by a sample of six houses.

n	Years (X)	Price (Y)
1	27	165
2	15	182
3	3	205
4	35	170
5	8	180
6	18	161

a. Find the least squares regression line $\hat{y} = b_0 + b_1 x$
b. Using the regression of house value on years, predict the value of a house that is 13 years old, rounded to the nearest dollar.
c. Find a 95% confidence interval and a 95% prediction interval for the result in part b.
d. Compute r and r^2 and explain what they mean.

2. Perform a quadratic regression on the data in problem 1 above. Does this improve the model? Justify your answer.

Chapter 6 – Non-Parametric Statistics

6.1 Chi-Square Tests

6.1.1 Goodness of Fit

To test whether a sample is from a particular distribution, we perform a goodness-of-fit test. Let's open a package of regular M&M's and count the number of each color. Suppose there are 15 brown M&M's, 13 yellow, 12 red, 32 blue 20 orange, and 16 green. First, we create a list of colors:

```
      COLORS←'Brown,Yellow,Red,Blue,Orange,Green'
```

Then we create a list of the corresponding counts:

```
      FREQ←15 13 12 32 20 16
```

We will test if the sample came from a uniform distribution; that is that the manufacturer produces the same number of each color:

```
   report uniform goodnessOfFit = COLORS FREQ
──────────────────────────────────────────────
  VALUE   OBS  EXP  ERR     CHISQ
  Brown    15   18   ¯3   0.5
  Yellow   13   18   ¯5   1.3889
  Red      12   18   ¯6   2
  Blue     32   18   14  10.889
  Orange   20   18    2   0.22222
  Green    16   18   ¯2   0.22222
  Total   108  108    0  15.222
  H₀: Uniform             H₁: not Uniform
 ┌───────────────────────┬───────────────────────┐
 │Test Statistic:        │P-Value:               │
 │χ² =15.222             │p=0.0094538            │
 ├───────────────────────┼───────────────────────┤
 │Critical Value:        │Significance Level:    │
 │χ²(α;df=5  )=11.070    │α=0.05                 │
 └───────────────────────┴───────────────────────┘
```

We reject the null hypothesis since the p-Value is less than 0.05 and the test statistic is greater than the critical value. It is evident that the colors are not uniformly distributed.

M&M/Mars used to publish the proportions of each color on the internet. They were 13% brown 14% yellow, 13% red, 24% blue, 20% orange, and 16% green. To test if the M&M's are still distributed this we say we perform a multinomial goodness of fit test. First, we set the proportions, making sure that they total 1:

```
      PROP←0.13 0.14 0.13 0.24 0.2 0.16
      +/PROP
1
```

Then we run the test and display the report showing a much better fit:

```
    report COLORS PROP multinomial goodnessOfFit = COLORS FREQ
```

```
  VALUE   OBS     EXP           ERR     CHISQ
  Brown    15   14.04   9.6000E¯1   0.065641
  Yellow   13   15.12  ¯2.1200E0    0.29725
  Red      12   14.04  ¯2.0400E0    0.29641
  Blue     32   25.92   6.0800E0    1.4262
  Orange   20   21.6   ¯1.6000E0    0.11852
  Green    16   17.28  ¯1.2800E0    0.094815
  Total   108  108     ¯3.5527E¯15  2.2988
  H₀: Multinomial         H₁: not Multinomial
 ┌──────────────────────┬────────────────────┐
 │Test Statistic:       │P-Value:            │
 │χ²=2.2988             │p=0.80644           │
 ├──────────────────────┼────────────────────┤
 │Critical Value:       │Significance Level: │
 │χ²(α;df=5  )=11.070   │α=0.05              │
 └──────────────────────┴────────────────────┘
```

Here we fail to reject the null hypothesis as the p-Value is large and the test statistic is less than the critical value.

We can also test data to see if data fit a particular discrete distribution. Does family size fit a Poisson distribution? TamStat will estimate the mean of the data and use that as the Poisson lambda parameter. The Family variable in the Student Database represents the number of siblings each student reported.

```
    .01 report poisson goodnessOfFit = #.SD.Family
```

```
  Value  Observed  Expected  Difference  ChiSquare
      0         2    7.0525    -5.0525    3.6196
      1        17   11.878      5.1222    2.2089
      2        11   10.002      0.99763   0.099502
    ≥ 3         8    9.0674    -1.0674    0.12564
  Total        38   38          0         6.0537

  Mean: 1.684210526

  H₀: Poisson  H₁: not Poisson
 ┌────────────────┬───────────────────┐
 │Test Statistic: │P-Value:           │
 │χ²=6.053664992  │p=0.04847          │
 ├────────────────┼───────────────────┤
 │Critical Value: │Significance Level:│
 │χ²(α;df=2)=9.210│α=0.01             │
 └────────────────┴───────────────────┘
  Conclusion:  Fail to reject H₀
```

In the above example, the data appear to be Poisson at the 1% significance level. Note that at the 5% level we would marginally reject the null hypothesis.

Other goodness-of-fit methods, such as those for continuous distributions, e.g. normal and exponential, use non-parametric methods which are discussed in Section 5.5.

6.1.2 Independence

A test of independence shows whether two qualitative variables are independent. From the student data, let’s see whether a person’s sex is independent of political party affiliation.

First let’s look at a contingency table:

```
    frequency #.SD.Sex #.SD.Party
    D  I   R
 F  3  2   4
 M  8  9  12
```

```
 report #.SD.Sex independent #.SD.Party

   *           D       I       R  Total

   F      3       2       4           9
          2.6053  2.6053  3.7895

   M      8       9       12         29
          8.3947  8.3947  12.211

   Total      11      11      16     38

  Key: Observed
       Expected

  H₀: Independent          H₁: not Independent

  Test Statistic:          P-Value:
  χ² =0.27795              p=0.87025

  Critical Value:          Significance Level:
  χ²(α;df=2  )=5.991       α=0.05

  Conclusion: Fail to reject H₀
```

Independence of variables is assumed unless there is compelling evidence to the contrary. The evidence in this case is insufficient to show a relationship between `Sex` and `Party`; therefore, those variables are assumed to be independent.

If the data are in summary form, just supply a right argument to `independent` and enter the count data in the grid:

```
      CTABLE←editTable 'F,M' 'D,I,R'

      report independent CTABLE
```

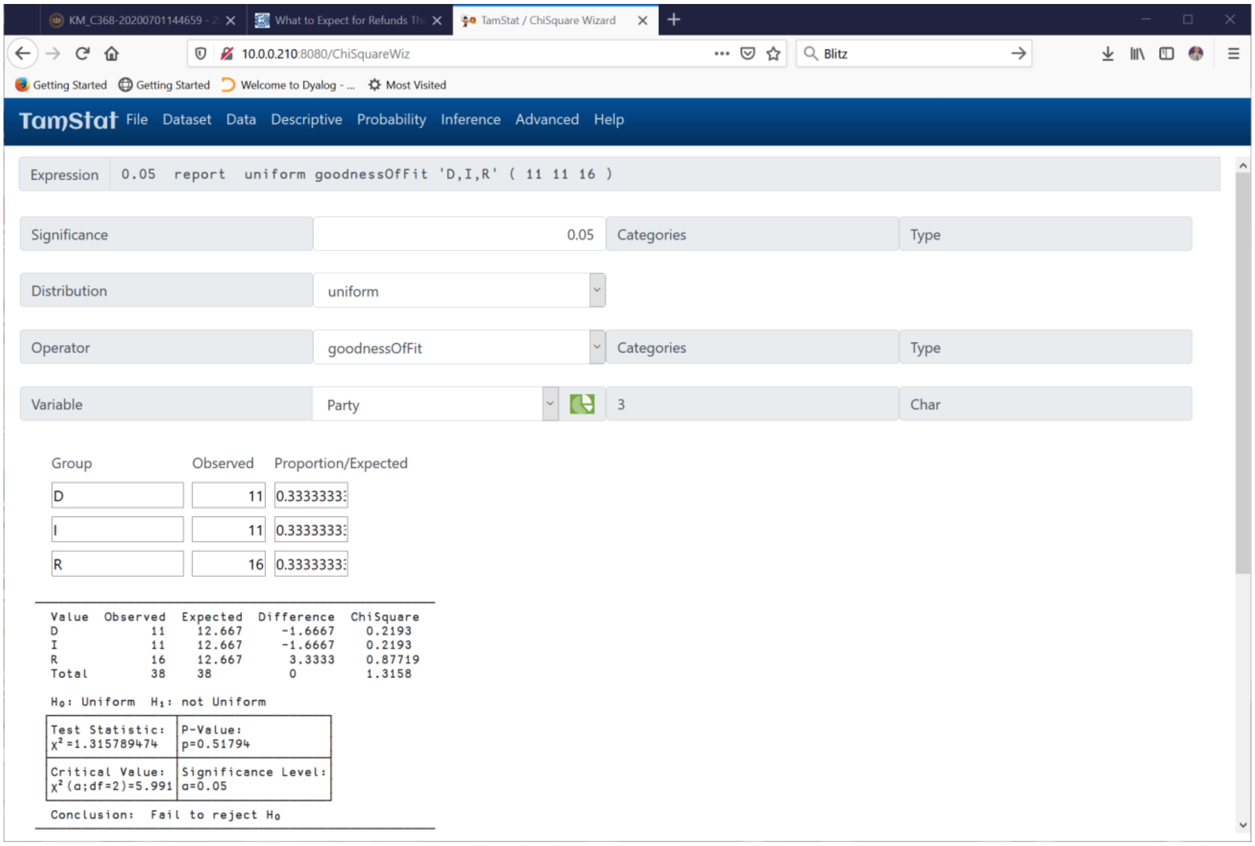
will create the same report as the expression:

```
      report #.SD.Sex independent #.SD.Party.
```

6.1.3 Chi-Square Test Wizard

To perform chi-square tests using the wizard, select “Advanced”, then select “Chi-Square Tests”. This screen allows the user to run goodness of fit and tests of independence. For goodness of fit, the user must select the appropriate distribution: uniform, multinomial or normal. “uniform” assumes that all the probabilities are equal, “multinomial” allows the user to select the probabilities. Both of these distributions work with categorical data. “normal” works with quantitative data and tests whether the data are suitably bell-shaped to come from a normal distribution.

Let’s perform a uniform goodness of fit test on student politics. We will first assume that there are an equal number of Democrats (D), Independents (I) and Republicans (R):



Expression `0.05 report uniform goodnessOfFit 'D,I,R' ( 11 11 16 )`

Significance	0.05	Categories	Type
Distribution	uniform		
Operator	goodnessOfFit	Categories	Type
Variable	Party	3	Char

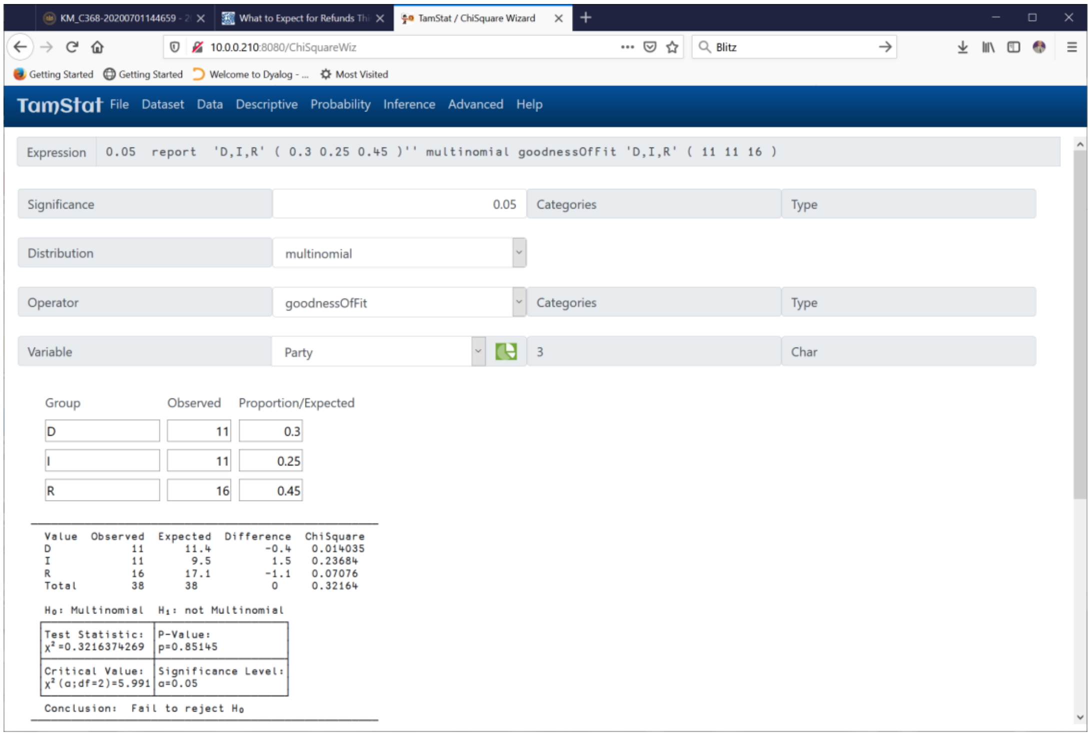
Group	Observed	Proportion/Expected
D	11	0.3333333:
I	11	0.3333333:
R	16	0.3333333:

```
Value  Observed  Expected  Difference  ChiSquare
D            11    12.667     -1.6667     0.2193
I            11    12.667     -1.6667     0.2193
R            16    12.667      3.3333     0.87719
Total        38    38          0          1.3158

H₀: Uniform  H₁: not Uniform
┌──────────────────┬────────────────────┐
│Test Statistic:   │P-Value:            │
│χ²=1.315789474    │p=0.51794           │
├──────────────────┼────────────────────┤
│Critical Value:   │Significance Level: │
│χ²(α;df=2)=5.991  │α=0.05              │
└──────────────────┴────────────────────┘
Conclusion:  Fail to reject H₀
```

Suppose the assumption is that the proportion of students differs by political parties. Let’s assume that in the Business School, 30% of the students are democrats, 25% are independents and 45% are republicans:

Party	Code	Probability
Democrat	D	25%
Independent	I	25%
Republican	R	5%
Total		100%

We now select “multinomial” for the distribution. The probability column now becomes editable. The initial values in this column will be divided equally. Simply replace the uniform probabilities with the new multinomial probabilities. To enter the assumed probabilities, select “Get Categories” and a grid showing each group with the number of observed values will appear.



Expression `0.05 report 'D,I,R' ( 0.3 0.25 0.45 )'' multinomial goodnessOfFit 'D,I,R' ( 11 11 16 )`

Significance: 0.05 | Categories | Type

Distribution: multinomial

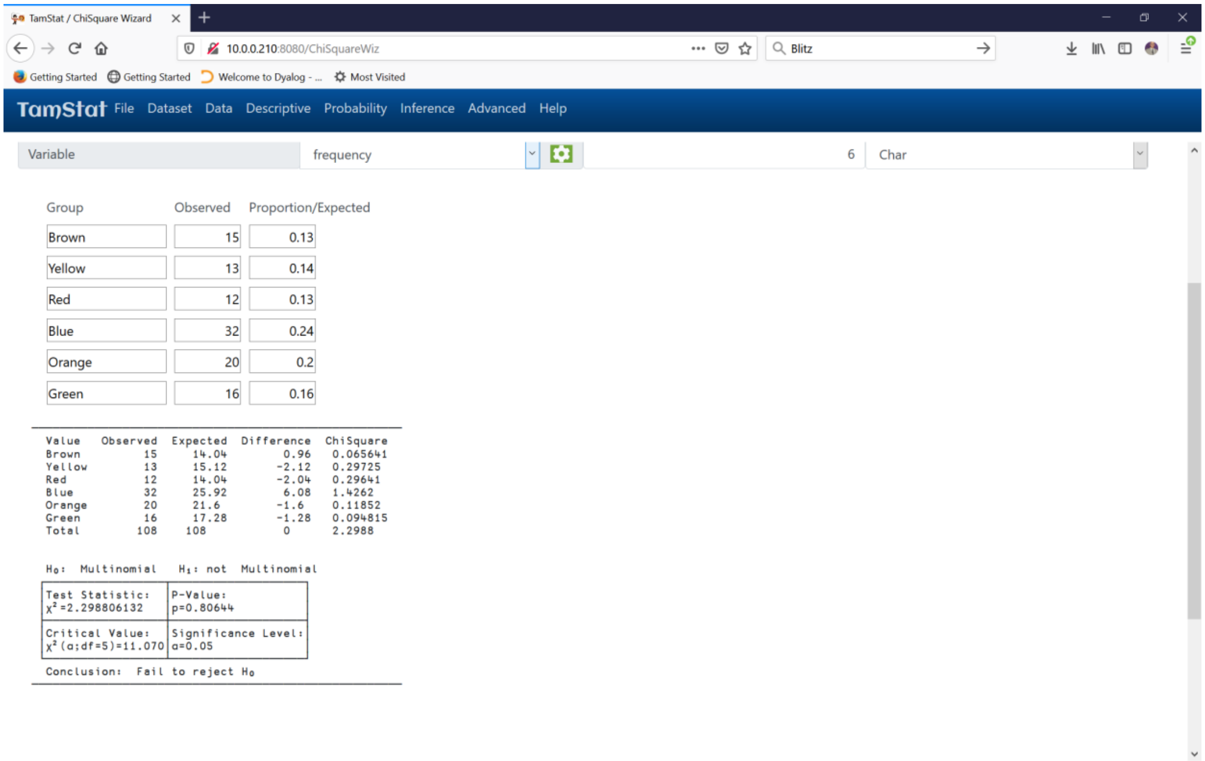
Operator: goodnessOfFit | Categories | Type

Variable: Party | 3 | Char

Group	Observed	Proportion/Expected
D	11	0.3
I	11	0.25
R	16	0.45

```
Value  Observed  Expected  Difference  ChiSquare
D            11      11.4        -0.4   0.014035
I            11       9.5         1.5   0.23684
R            16      17.1        -1.1   0.07076
Total        38      38           0     0.32164

H₀: Multinomial  H₁: not Multinomial
┌────────────────┬─────────────────┐
│Test Statistic: │P-Value:         │
│χ²=0.3216374269 │p=0.85145        │
├────────────────┼─────────────────┤
│Critical Value: │Significance Level:
│χ²(α;df=2)=5.991│α=0.05           │
└────────────────┴─────────────────┘
Conclusion:  Fail to reject H₀
```

In some cases, it is not practical to have a database and we must use summary data. For example, in counting the number of M&M's we don't have a record for each M&M! So, we select "frequency" for variable then select 6 for the number of Categories. In the "Group" column, enter the colors, in the "Observed" column, enter the number of M&M's of each color, and in the "Proportion/Expected" Column, enter the proportions (or percentages) in each category:

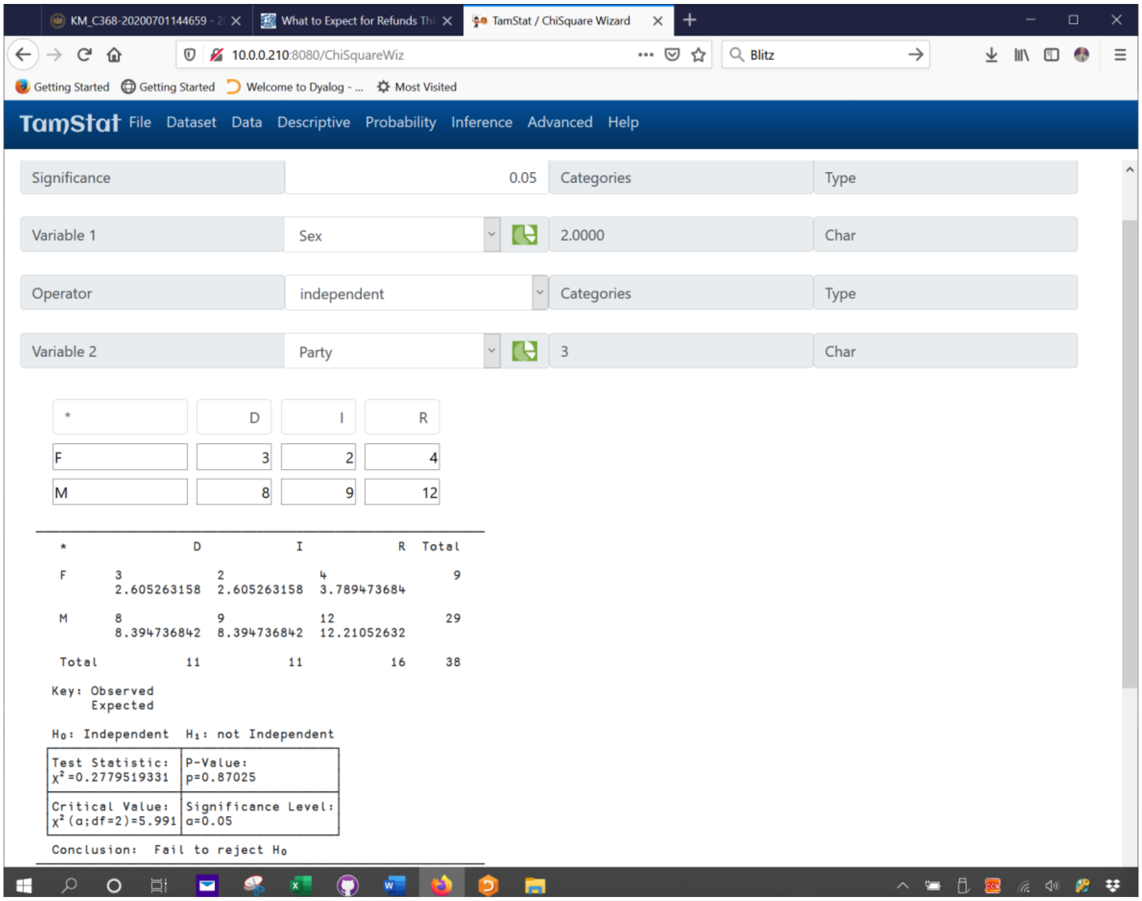


Variable: frequency | 6 | Char

Group	Observed	Proportion/Expected
Brown	15	0.13
Yellow	13	0.14
Red	12	0.13
Blue	32	0.24
Orange	20	0.2
Green	16	0.16

```
Value    Observed  Expected  Difference  ChiSquare
Brown          15     14.04        0.96   0.065641
Yellow         13     15.12       -2.12   0.29725
Red            12     14.04       -2.04   0.29641
Blue           32     25.92        6.08   1.4262
Orange         20     21.6        -1.6    0.11852
Green          16     17.28       -1.28   0.094815
Total         108    108           0      2.2988

H₀:  Multinomial   H₁: not  Multinomial
┌─────────────────┬──────────────────┐
│Test Statistic:  │P-Value:          │
│χ²=2.298806132   │p=0.80644         │
├─────────────────┼──────────────────┤
│Critical Value:  │Significance Level:
│χ²(α;df=5)=11.070│α=0.05            │
└─────────────────┴──────────────────┘
Conclusion:  Fail to reject H₀
```

To perform a test of independence, select the “independent” operator; then select two nominal variables from the database. For example, choose “Sex” and “Party”; TamStat will populate the contingency table, then calculate expected values and perform a test of independence:



Significance	0.05	Categories	Type
Variable 1	Sex	2.0000	Char
Operator	independent	Categories	Type
Variable 2	Party	3	Char

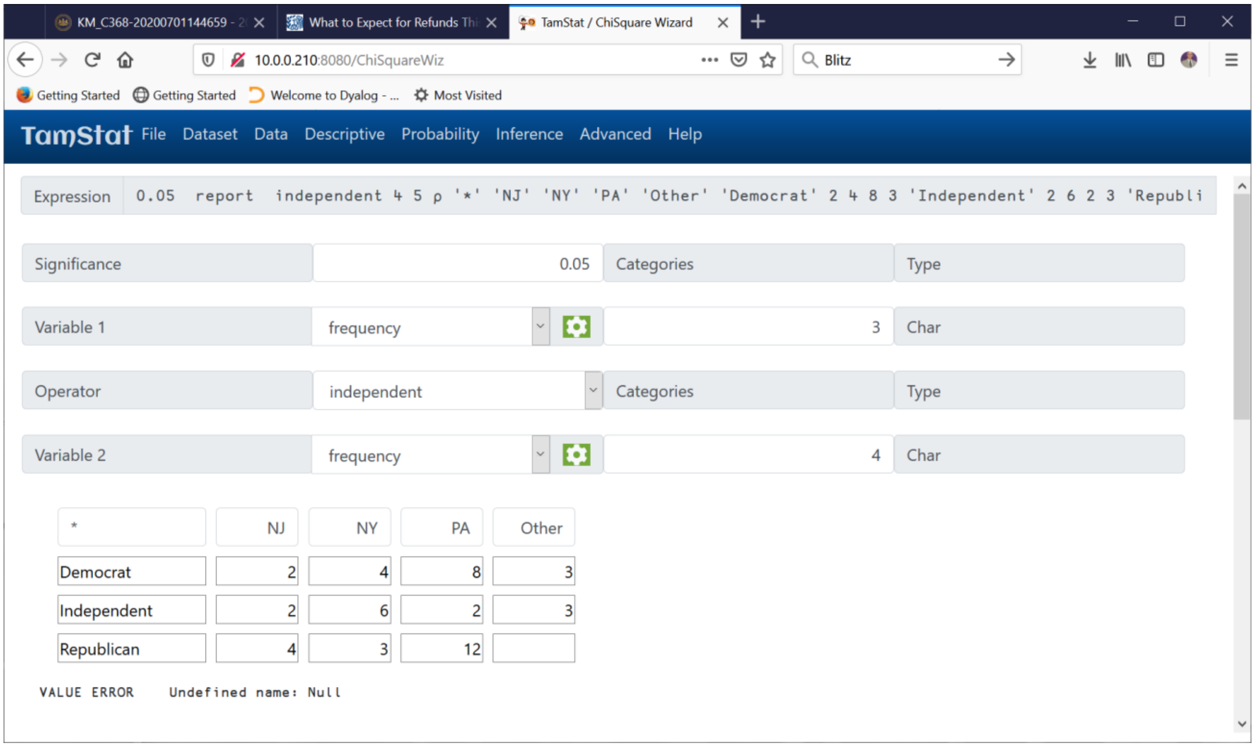
*	D	I	R
F	3	2	4
M	8	9	12

```
 *               D            I             R  Total

 F      3            2            4                9
        2.605263158  2.605263158  3.789473684

 M      8            9            12              29
        8.394736842  8.394736842  12.21052632

 Total          11           11            16     38

Key: Observed
     Expected

H₀: Independent  H₁: not Independent
┌────────────────┬──────────────────┐
│Test Statistic: │P-Value:          │
│χ²=0.2779519331 │p=0.87025         │
├────────────────┼──────────────────┤
│Critical Value: │Significance Level:│
│χ²(α;df=2)=5.991│α=0.05            │
└────────────────┴──────────────────┘
Conclusion:  Fail to reject H₀
```

To work with summary data, you need to create a contingency table. Suppose we want to see if `Party` is independent of `State`. We have the following summary information:

Party	**New Jersey**	**New York**	**Pennsylvania**	**Other**	**Total**
Democrat	2	4	8	3	17
Independent	2	6	2	3	13
Republican	4	3	12	3	22
Total	8	13	22	9	52

To create this table, go to “`Advanced`”, then “`ChiSquareTests`”. In the operator field, select “independent”, then select “frequency” for “Variable 1” and “Variable 2”. Select the number of categories for each variable, i.e. 3 for Variable 1 and 4 for category 2; this will resize the editable contingency table. Enter the Row and Column Headings; then populate the table with quantities. Do not enter the row and column totals.



TamStat File Dataset Data Descriptive Probability Inference Advanced Help

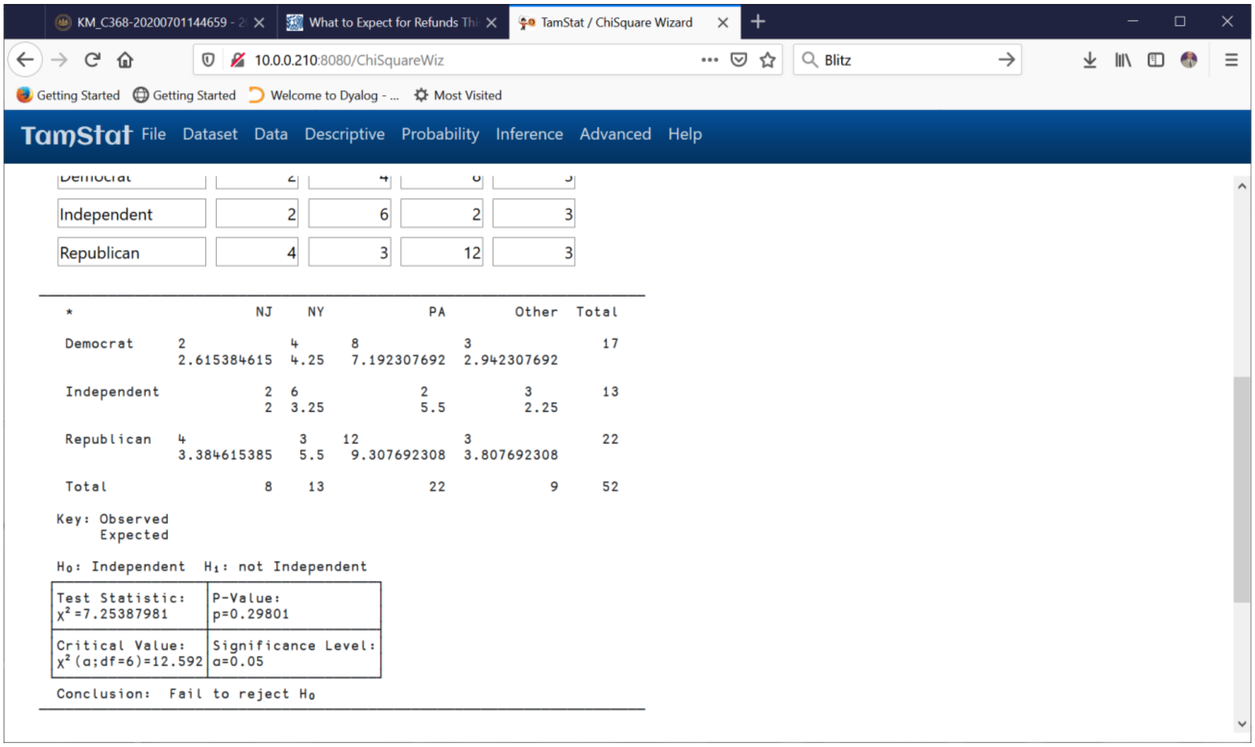
```
Expression  0.05 report independent 4 5 ρ '*' 'NJ' 'NY' 'PA' 'Other' 'Democrat' 2 4 8 3 'Independent' 2 6 2 3 'Republi
```

Significance	0.05	Categories	Type
Variable 1	frequency	3	Char
Operator	independent	Categories	Type
Variable 2	frequency	4	Char

*	NJ	NY	PA	Other
Democrat	2	4	8	3
Independent	2	6	2	3
Republican	4	3	12	

```
VALUE ERROR    Undefined name: Null
```

When the contingency table is completely populated, TamStat will perform the test. Scroll down to see the results:



TamStat File Dataset Data Descriptive Probability Inference Advanced Help

Independent	2	6	2	3
Republican	4	3	12	3

```
  *                     NJ     NY              PA          Other   Total

  Democrat     2               4      8                3              17
               2.615384615  4.25   7.192307692   2.942307692

  Independent             2  6                  2          3          13
                          2  3.25             5.5          2.25

  Republican   4              3    12                3                 22
               3.384615385    5.5  9.307692308   3.807692308

  Total                    8    13             22              9      52

 Key: Observed
      Expected

 H₀: Independent  H₁: not Independent
┌────────────────┬──────────────────┐
│Test Statistic: │P-Value:          │
│χ² =7.25387981  │p=0.29801         │
├────────────────┼──────────────────┤
│Critical Value: │Significance Level:│
│χ² (α;df=6)=12.592│α=0.05          │
└────────────────┴──────────────────┘
 Conclusion:  Fail to reject H₀
```

The p-Value 0.298 is greater than the significance level and the chi-square statistic is less than the critical value, so we fail to reject the null hypothesis and conclude that party affiliation is independent of state of residence.

When the shape of the distribution is unknown, and there are no assumptions about normality, we are left with ordinal data. Non-parametric statistics are useful for dealing with ordinal data. In particular, we are dealing with hypothesis tests for the median rather than the mean.

For small-sample cases, we will reject the null hypothesis when the test statistic is smaller than the critical value because the distribution is left-tailed. In large-sample cases, we can use the usual right-tailed test because of the symmetry of the normal distribution.

6.2 Testing the median – Sign Test

A computer manufacturer claims that the median time to failure for their hard drives is 14,400 hours. Test the claim at the 1% level of significance. The following vector lists the times to failure for a sample of 16 hard drives:

```
Time2Failure←330 620 1870 2410 4620 6396 7822 8102 8309 12882 14419 16092 18384 20916 23812 25814
```

The sign test simply counts the number of values above and below the hypothesized median. There are 10 values below 14,400 and six above. To perform the sign test for a sample from a single population, use the hypothesis operator with **median** as the left operand:

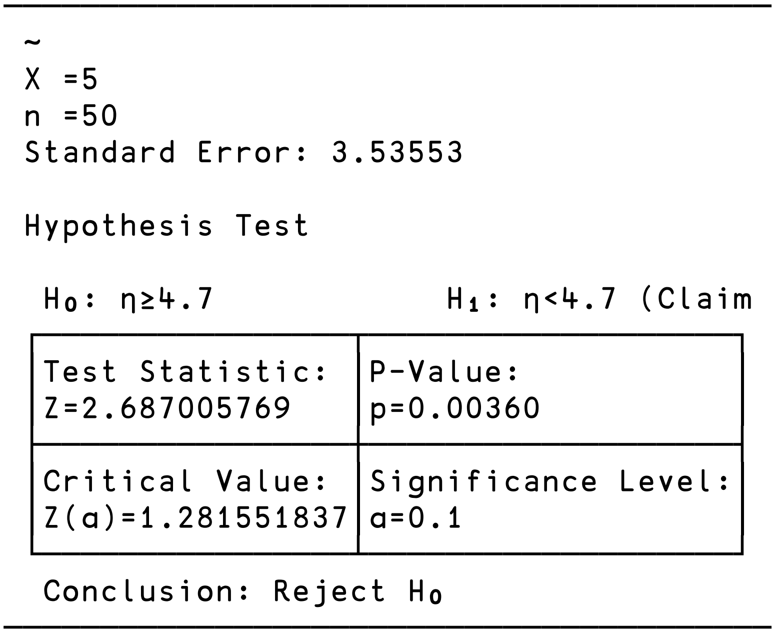
```
    .01 report Time2Failure median hypothesis < 14400
```

```
 ~
 X =8205.50000
 n =16
 Standard Error: 2.00000

 Hypothesis Test

  H₀: η≥14400         H₁: η<14400 (Claim)
 ┌───────────────────┬──────────────────────┐
 │Test Statistic:    │P-Value:              │
 │X=6                │p=0.22725             │
 ├───────────────────┼──────────────────────┤
 │Critical Value:    │Significance Level:   │
 │X(α;N=16)=2        │α=0.01                │
 └───────────────────┴──────────────────────┘
  Conclusion: Fail to reject H₀
```

6.2.1 Testing the median – Large Sample

The national median for hospital stats is 4.7 days. At a local hospital the lengths of stays for 35 out of 50 patients was less than 4.7 days. At 10% significance, is median stay at the local hospital less than the national median?

```
   HospitalStays←35 15/4 5
   0.1 report HospitalStays median hypothesis < 4.7
```



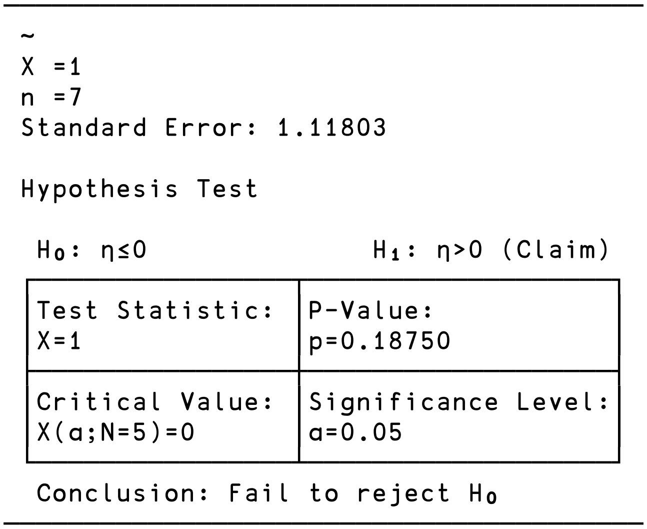
```
~
X =5
n =50
Standard Error: 3.53553

Hypothesis Test

 H₀: η≥4.7            H₁: η<4.7 (Claim
┌────────────────┬──────────────────┐
│Test Statistic: │P-Value:          │
│Z=2.687005769   │p=0.00360         │
├────────────────┼──────────────────┤
│Critical Value: │Significance Level:│
│Z(α)=1.281551837│α=0.1             │
└────────────────┴──────────────────┘
 Conclusion: Reject H₀
```

Observe that the test statistic is normal for large samples and in this case it exceeds the critical value resulting in the rejection of the null hypothesis.

6.2.2 Sign Test on Paired Data

We can also use the sign test on paired data. We simply take the differences between each pair of data and perform the sign test on the difference. A supermarket wants to test out a new cash register. Seven cashiers are selected randomly, and each uses the old register for three minutes, followed by the new register for the minutes. The number of items processed during each three-minute period are recorded.

```
    OldCR←60 70 55 75 62 52 58 ⍝ Seven cashiers using old register
    NewCR←65 71 55 75 65 57 57 ⍝ The same seven cashiers using new register

    report (NewCR-OldCR) median hypothesis > 0
```



```
~
X =1
n =7
Standard Error: 1.11803

Hypothesis Test

 H₀: η≤0              H₁: η>0 (Claim)
┌──────────────┬──────────────────┐
│Test Statistic:│P-Value:          │
│X=1           │p=0.18750         │
├──────────────┼──────────────────┤
│Critical Value:│Significance Level:│
│X(α;N=5)=0    │α=0.05            │
└──────────────┴──────────────────┘
 Conclusion: Fail to reject H₀
```

6.3 The Wilcoxon Signed-Rank Test

This is also a paired-data test. It is slightly better than the Sign test for paired data because it accounts for the magnitude of the data indirectly in the form of ranking. We can perform this test using the same data in the previous section. TamStat will do the signed-rank test when the **paired** operator is applied to the median:

```
     report OldCR median paired hypothesis > NewCR
```

```
 ~
 X =¯1
 s =0
 n =7

 Hypothesis Test

  H0: ηd≤0              H1: ηd>0 (Claim)
 ┌─────────────────┬────────────────────┐
 │Test Statistic:  │P-Value:            │
 │T=1.5            │p=0.05000           │
 ├─────────────────┼────────────────────┤
 │Critical Value:  │Significance Level: │
 │T(α)=0           │α=0.05              │
 └─────────────────┴────────────────────┘
  Conclusion: Fail to reject H0
```

6.4 The Mann-Whitney U Test

This is a two-population test comparing the medians of two populations of unknown shapes. Since the data are not paired, the sample sizes do not have to be the same. This test is like the two-sample t-test except that we do not make any assumptions about the shape of the distributions of each population. Instead of comparing the means of two populations, we will be comparing the medians of those populations.

When the samples are small ($n_1, n_2 \leq 10$), the test statistic is distributed according to the mannWhitneyU distribution; for large samples, the test statistic is approximately normal.

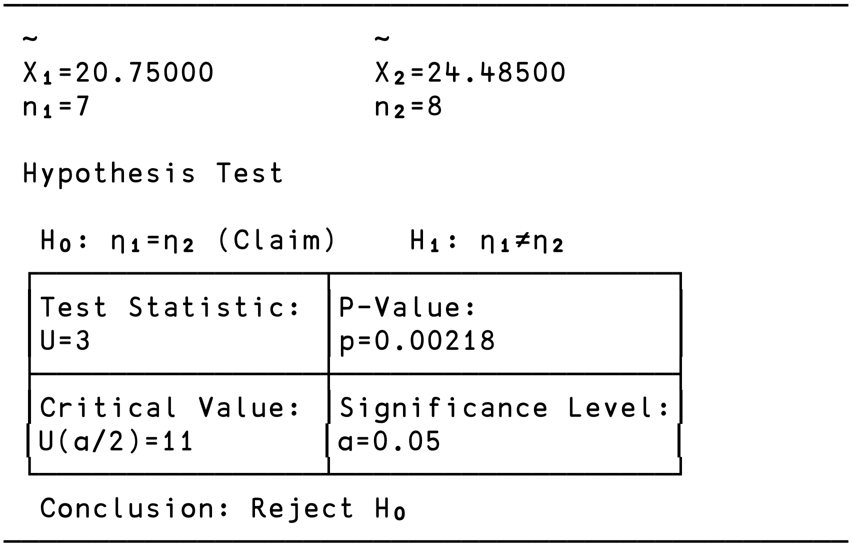
6.4.1 The Mann-Whitney U Test – Small Samples

Is there a difference between hourly compensation for healthcare and education workers? A sample of 7 healthcare payrates and 8 education payrates are compared:

```
    Health←20.1 19.8 22.36 18.75 21.9 22.96 20.75

     Education←26.19 23.88 25.5 21.64 24.85 25.3 24.12 23.45
```

Since both sample sizes are less than 10, TamStat will use the **mannWhitneyU** distribution:

```
      report Health median hypothesis = Education
```

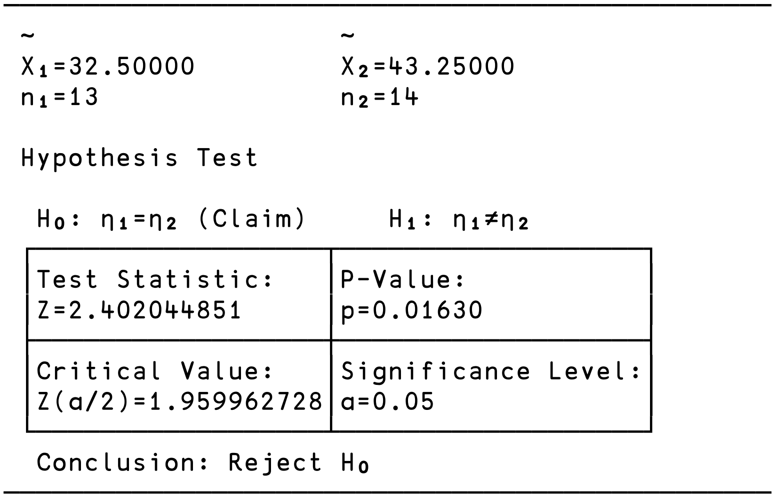


```
~                  ~
X₁=20.75000        X₂=24.48500
n₁=7               n₂=8

Hypothesis Test

 H₀: η₁=η₂ (Claim)    H₁: η₁≠η₂
┌────────────────┬───────────────────┐
│Test Statistic: │P-Value:           │
│U=3             │p=0.00218          │
├────────────────┼───────────────────┤
│Critical Value: │Significance Level:│
│U(α/2)=11       │α=0.05             │
└────────────────┴───────────────────┘
 Conclusion: Reject H₀
```

6.4.2 The Mann-Whitney U Test – Large Samples

Is there a difference in incomes for PBS viewers and non-PBS viewers? Annual salaries in thousands are collected from and random sample of 14 PBS viewers and 13 non-PBS viewers.

```
PBS←24.5 39.4 36.8 43 57.96 32 61 34 43.5 55 39 62.5 61.4 53

NonPBS←41 32.5 33 21 40.5 32.4 16 21.5 39.5 27.6 43.5 51.9 27.8

report PBS median hypothesis = NonPBS
```



```
~                  ~
X₁=32.50000        X₂=43.25000
n₁=13              n₂=14

Hypothesis Test

 H₀: η₁=η₂ (Claim)     H₁: η₁≠η₂
┌───────────────────┬───────────────────┐
│Test Statistic:    │P-Value:           │
│Z=2.402044851      │p=0.01630          │
├───────────────────┼───────────────────┤
│Critical Value:    │Significance Level:│
│Z(α/2)=1.959962728 │α=0.05             │
└───────────────────┴───────────────────┘
 Conclusion: Reject H₀
```

Notice that TamStat uses the normal distribution for the test statistic since the individual sample sizes are greater than 10.

6.5 The Kruskal-Wallis Test – Three or more groups

The Kruskal-Wallis Test compares the medians of three or more groups. It is comparable to one-way ANOVA for means.

Three brands of wood stains are compared. The number of months until the wood needed to be retreated was collected for various samples.

```
      Brand1←124 75 69 70 122 73

      Brand2←39 29 26 28 26

      Brand3←25 14 26 35 34 16 12

      report median hypothesis = Brand1 Brand2 Brand3
```

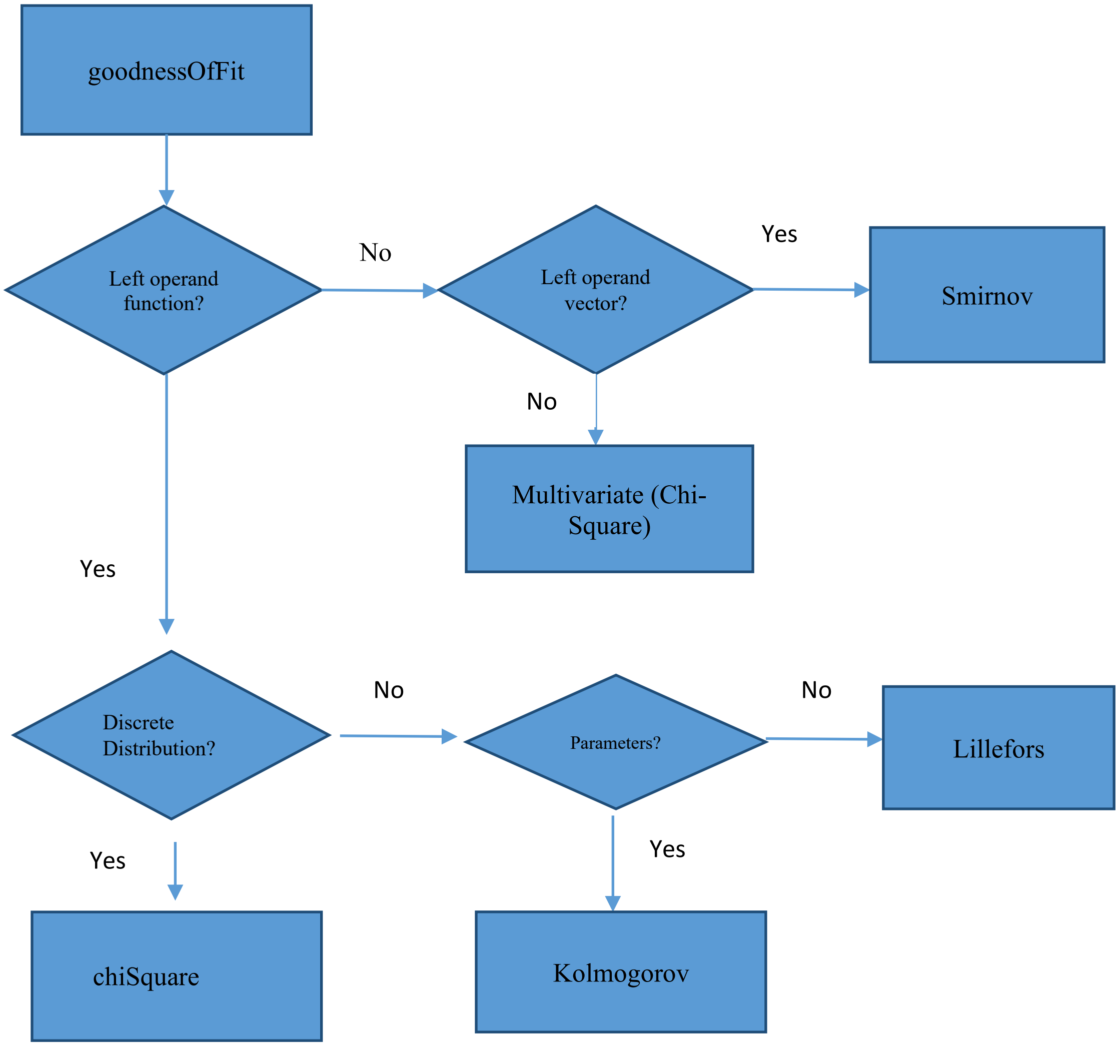
```
 Kruskal-Wallis Test
  ~      ~      ~
  X₁=74  X₂=28  X₃=25
  n₁=6   n₂=5   n₃=7

 H₀: η₁=η₂=η₃
 H₁: At least one median is different.
┌─────────────────┬──────────────────────┐
│Test Statistic:  │P-Value:              │
│χ²=12.23759398   │p=0.00220             │
├─────────────────┼──────────────────────┤
│Critical Value:  │Significance Level:   │
│χ²(α;df=2)=5.991 │α=0.05                │
└─────────────────┴──────────────────────┘
 Conclusion:  Reject H₀
```

The test statistic has a Chi-Square distribution with $k - 1$ degrees of freedom where k is the number of groups.

6.6 Goodness-of-Fit Tests

Non-parametric tests are available to test whether a sample comes from a particular statistical distribution or family of distributions, or whether two samples come from the same distribution. These tests include the Kolmogorov Test for testing continuous distributions, the Lillefors test for testing for families of distributions, and the Smirnov Test for testing two distributions. The chi-Square test for discrete distributions is described in Section 4.1. The **goodnessOfFit** operator will determine which type of test to perform based on the characteristics of its arguments and operands. The decision tree for the goodnessOfFit operator is illustrated on the next page.

The **goodnessOfFit** operator takes a distribution function as its left operand, and a relational function as its right operand. The right argument is a vector containing the sample data. The left argument is optional; if it is specified, then the test is whether the data come from that distribution. Otherwise, the test is whether the data come from a family of distributions. If the left operand is a vector instead of a function, then it is considered a second sample to be compared with the right argument.



Decision tree for the `goodnessOfFit` Operator

6.6.1 The Kolmogorov Test

The Kolmogorov test works best with continuous distributions. The test works for small samples as well as large samples. Both one-sided and two-sided tests are available.

Five children were selected from a class at random and timed in a short race. The times were 6.3, 4.2 4.7, 6 and 5.7 seconds. The previous race times were uniformly distributed between 4 and 8 seconds. Test whether the race time distribution has improved.

The report on the next page shows that there has not been a significant improvement in race times, and that the current race times are still distributed uniformly between 4 and 8 seconds. Note that we use the “rectangular” distribution in TamStat which is the continuous analog of the discrete uniform distribution. Also note that the largest positive and largest negative differences are flagged in the report.

```
      report  4 8 rectangular goodnessOfFit <  6.3 4.2 4.7 6 5.7
──────────────────────────────────────────
 kolmogorov Test
 i  Value  S(x)  F(x)      T+     T-
 1    4.2   0.2  0.05    0.05   0.15  *
 2    4.7   0.4  0.175  -0.025  0.225
 3    5.7   0.6  0.425   0.025  0.175
 4    6     0.8  0.5    -0.1    0.3
 5    6.3   1    0.575  -0.225  0.425 *

   Mean: 5.38    Sample Size: N =   5

  H₀:F(x)≥F*(x)  H₁:F(x)<F*(x)
 ┌─────────────────┬─────────────────────┐
 │Test Statistic:  │P-Value:             │
 │T=0.425          │p=0.12367            │
 ├─────────────────┼─────────────────────┤
 │Critical Value:  │Significance Level:  │
 │T(α)=0.509       │α=0.05               │
 └─────────────────┴─────────────────────┘
  Conclusion:  Fail to reject H₀
──────────────────────────────────────────
```

Automobile emissions from the previous year have been measured and were normally distributed with a mean of 5.6 and a standard deviation of 1.2. Twelve cars were randomly selected, and the following emissions measurements taken:

4.8, 6.2, 6.0, 5.9, 6.6, 5.5, 5.8, 5.9, 6.3, 6.6, 6.2, 5.0

Do the current emissions have the same distribution as the previous year?

```
     X←4.8 6.2 6 5.9 6.6 5.5 5.8 5.9 6.3 6.6 6.2 5

     report 5.6 1.2 normal goodnessOfFit = X
──────────────────────────────────────────────────────
 kolmogorov Test
  i  Value      S(x)      F(x)         T+         T-
  1    4.8  0.083333  0.25249   0.25249   -0.16916
  2    5    0.16667   0.30854   0.2252    -0.14187
  3    5.5  0.25      0.46679   0.30013   -0.21679
  4    5.8  0.33333   0.56618   0.31618   -0.23285  *
  5    5.9  0.5       0.59871   0.18204   -0.098706
  6    5.9  0.5       0.59871   0.18204   -0.098706
  7    6    0.58333   0.63056   0.13056   -0.047225
  8    6.2  0.75      0.69146   0.024796   0.058537
  9    6.2  0.75      0.69146   0.024796   0.058537
 10    6.3  0.83333   0.72017  -0.029834   0.11317
 11    6.6  1         0.79767  -0.11899    0.20233  *
 12    6.6  1         0.79767  -0.11899    0.20233

   Mean: 5.9    Sample Size: N =   12

  H₀:F(x)=F*(x)  H₁:F(x)≠F*(x)
 ┌────────────────┬────────────────────┐
 │Test Statistic: │P-Value:            │
 │T=0.3161839595  │p=0.14499           │
 ├────────────────┼────────────────────┤
 │Critical Value: │Significance Level: │
 │T(α)=0.375      │α=0.05              │
 └────────────────┴────────────────────┘
  Conclusion:  Fail to reject H₀
──────────────────────────────────────────────────────
```

6.6.2 The Lillefors Test

When the exact distribution is unknown, one can test whether the data come from a family of distributions. For example, if the data appear bell-shaped, we can test whether the sample comes from a normal distribution. The student survey contains the weights of students. Let us test whether the student weights are normally distributed:

```
report normal goodnessOfFit = #.SD.Weight
```

```
 Lillefors Test
  i    Xi        Zi     S(Zi)     F(Zi)        T+          T-
  1 100    -1.6545   0.026316  0.049017   0.049017  -0.022702
  2 105    -1.5359   0.052632  0.062281   0.035965  -0.0096497
  3 115    -1.2988   0.078947  0.097008   0.044376  -0.01806
  4 115    -1.2988   0.10526   0.097008   0.01806    0.0082556
  5 120    -1.1802   0.13158   0.11895    0.013689   0.012626
 ................................................................
  8 139.5  -0.71788  0.21053   0.23642    0.052206  -0.02589   *
 ................................................................
 22 165    -0.11325  0.57895   0.45492   -0.097716   0.12403   *
 ................................................................
 34 220     1.1908   0.89474   0.88314    0.014722   0.011594
 35 225     1.3094   0.92105   0.9048     0.010064   0.016252
 36 245     1.7836   0.94737   0.96276    0.041705  -0.015389
 37 260     2.1393   0.97368   0.98379    0.036425  -0.010109
 38 280     2.6135   1         0.99552    0.021835   0.0044811

   Mean: 169.7763158  Standard Deviation: 42.17478069    Sample Size: N =   38

  H0:normal  H1:not normal
 ┌─────────────────┬────────────────────┐
 │Test Statistic:  │P-Value:            │
 │T=0.1240315761   │p=0.16469           │
 ├─────────────────┼────────────────────┤
 │Critical Value:  │Significance Level: │
 │T(α)=0.156       │α=0.05              │
 └─────────────────┴────────────────────┘
  Conclusion:  Fail to reject H0
```

Note when the number of observations is large, the report only displays the first and last 5 observations as well as the observations containing the largest and smallest differences.

6.6.3 The Smirnov Test

A random sample of 9 packages from a delivery service is taken and each parcel is weighed. A random sample of 12 packages from another delivery service is taken and those packages are also weighed. Are the distributions of weights for each delivery service the same?

```
 ⍝ Weights for Delivery Service X
X←7.6 8.4 8.6 8.7 9.3 9.9 10.1 10.6 11.2

⍝ Weights for Delivery Service Y
Y←5.2 5.7 5.9 6.5 6.8 8.2 9.1 9.8 10.8 11.3 11.5 12.3 12.5 13.4 14.6
```

```
   report X goodnessOfFit = Y
```

```
smirnov Test
 i    X    Y  S1        S2     S1-S2
 1  0    5.2   0  0.066667  0.066667
 2  0    5.7   0  0.13333   0.13333
 3  0    5.9   0  0.2       0.2
 4  0    6.5   0  0.26667   0.26667
 5  0    6.8   0  0.33333   0.33333
.....................................
18 11.2  0     1  0.6       0.4      *
.....................................
20  0   11.5   1  0.73333   0.26667
.....................................
21  0   12.3   1  0.8       0.2
22  0   12.5   1  0.86667   0.13333
23  0   13.4   1  0.93333   0.066667
24  0   14.6   1  1         0        *
24  0   14.6   1  1         0        *

Sample Size: N =  9 M = 15

 H0:F(x)=G(x)  H1:F(x)≠G(x)
┌─────────────────┬────────────────────┐
│Test Statistic:  │P-Value:            │
│T=0.4            │p=0.33060           │
├─────────────────┼────────────────────┤
│Critical Value:  │Significance Level: │
│T(α)=0.573       │α=0.05              │
└─────────────────┴────────────────────┘
 Conclusion:  Fail to reject H0
```

It appears that the distribution of weights for the delivery services are the same.

6.7 Exercises

2. You ask 100 persons what color of car each prefers. The table below lists the observed frequencies for such a survey. Test the hypothesis that 10% of all persons like black cars, 25% like blue cars, 20% like red cars, 10% like white cars, and 35% like other colors.

Color	Black	Blue	Red	White	Other
Observed	21	29	14	6	30

3. A survey of students from a regional university indicates the following counts:

	Home State		
Party	New Jersey	New York	Pennsylvania
Democrat	8	6	11
Independent	7	9	5
Republican	15	12	24

Is party affiliation independent of home state? Test at 10% significance.

Chapter 7 – Simulation

7.1 Apartment Rental Problem

You have an apartment complex with 40 units. Each unit rents for $500 a month. Demand follows a discrete uniform distribution between 30 and 40. Your monthly expenses average $15,000 a month with a standard deviation of $3,000. What is your expected profit? What is the standard deviation? What is the probability that you will lose money?

```
      RENTED←30 40 uniform randomVariable 1000
      EXPENSES←15000 3000 normal randomVariable 1000
      PROFIT←(500 times RENTED) - EXPENSES
      mean PROFIT
2483.6
      sdev PROFIT
3427
      proportion PROFIT<0
0.236
```

7.2 Newsvendor Problem

You sell fruit in the marketplace. You can buy apples wholesale for 50 cents each if you buy in multiples of 10. You can sell individual apples for 75 cents each. If the apples don't sell that day, you must discard them. Your daily demand averages 37 apples and follows a Poisson distribution. Should you buy 30 apples or 40 apples?

This is an example of the classical newsvendor problem:

$$\pi = E\left[p \min(q, D)\right] - cq$$

where π = profit, p = unit price, q = quantity ordered, c = unit cost, and D = demand.

```
      PRICE ← 0.75     ⍝ Unit Price
      COST ← 0.50      ⍝ Wholesale cost
      DEMAND ← 37 poisson randomVariable 1000
      QUAN←30          ⍝ Quantity ordered = 30
      REVENUE ← PRICE times QUAN min DEMAND
      EXPENSE ← COST times QUAN
      PROFIT ← REVENUE - EXPENSE
      mean PROFIT      ⍝ Expected profit = $7.19 if you buy30 apples
7.1872
      sdev PROFIT      ⍝ Standard deviation is 99 cents.
0.98842
      proportion PROFIT < 0  ⍝ Virtually no chance of losing money
0
      QUAN ← 40          ⍝ Quantity ordered = 40
      REVENUE ← PRICE times QUAN min DEMAND
      EXPENSE ← COST times QUAN
      PROFIT ← REVENUE-EXPENSE
      mean PROFIT     ⍝ Expected profit = $6.57
6.571
      sdev PROFIT     ⍝ Standard deviation is $3.46
```

```
3.4611
      proportion PROFIT < 0 ⍝ 4.5% chance of losing money
0.045
```

Obviously, it is more profitable and less risky to buy 30 apples, than to buy 40 apples.

7.3 Monte Hall Problem

In the game show “Let’s Make a Deal”, Monte Hall shows three doors. Behind one of the doors is a brand-new car. Behind the other two doors is a goat. Suppose the contestant selects Door #1. Then Monte Hall selects Door #2 and asks the contestant if he would like to switch to Door #3. Should the contestant keep his original choice or switch to Door #3? Does it matter?

Let’s generate 500 random samples where the car is equally likely to be behind any one of the three doors.

```
     CAR←3 uniform randomVariable 500
```

If the car is actually behind door #1, Monte will select Door #2 or Door #3 with equal probability:

```
     MONTE←2 3 uniform randomVariable 500
```

If the car is actually behind door #2, Monte will select Door #3; If the car is actually behind door #3, Monte will select Door #2:

```
     ADD ← ¯1 × MONTE = 3
     MONTE ← MONTE + ADD × CAR = MONTE
```

Probability that car is behind Door #1

```
      proportion CAR = 1
0.36
```

If you switch, you pick the door that Monte doesn’t open. Then if Monte picks door 2, you pick door 3, or if Monte picks door 3, you pick door 2:

```
      SWITCH ← 5 - MONTE
```

Probability you select the car if you switch:

```
       proportion SWITCH = CAR
0.64
```

The probability is only about 1/3 if you remain with your original door, but increases to 2/3 if you switch!

7.4 Auto Insurance Problem

Assuming you won the car on Let’s Make A Deal, you now have to insure it. The Insurance company wants to determine how much money it need to have on hand to pay claims. The probability distribution of weekly claims is:

Claims	1	2	3	4	5	6	7	8	9
Prob	0.05	0.06	0.10	0.17	0.28	0.14	0.08	0.07	0.05

The average repair bill $1700 with a standard deviation of $400. The company also pays claims for “totaled” cars; the probability of these claims is 20% per week. Totalled claims range from $2,000 to $35,000; the most likely claim is $13,000.

To generate a random variable for a custom distribution, we use the multinomial distribution. First we list the possible number of claims each month:

```
    X←1 2 3 4 5 6 7 8 9
    P←0.05 0.06 0.1 0.17 0.28 0.14 0.08 0.07 0.05
```

Let’s generate 1000 iterations:

```
    N←1000
```

Now we’ll generate the weekly repair claims:

```
      REPAIR_CLAIMS←X P multinomial randomVariable N
```

The repair amounts follow a normal distribution:

```
      REPAIR_AMTS←1700 400 normal randomVariable N
```

To generate totaled claims, we use the binomial distribution with n = 1 and π = 0.2.

```
      TOTALED_CLAIMS←1 0.2 binomial randomVariable N
```

When we know the limits and mode of the distribution, but not the shape, we can use the triangular distribution to model amount of the totaled claims:

```
      TOTALED_AMT←2000 13000 35000 triangular randomVariable N
```

The total payout each week is the sum of the repair claims and the totaled claims:

```
      PAYOUT←(REPAIR_CLAIMS × REPAIR_AMTS)+(TOTALED_CLAIMS × TOTALED_AMT)
```

The average weekly payout is about $10,000.

```
     mean PAYOUT
11564
```

But the standard deviation is about $8,000.

```
    sdev PAYOUT
8221.4
```

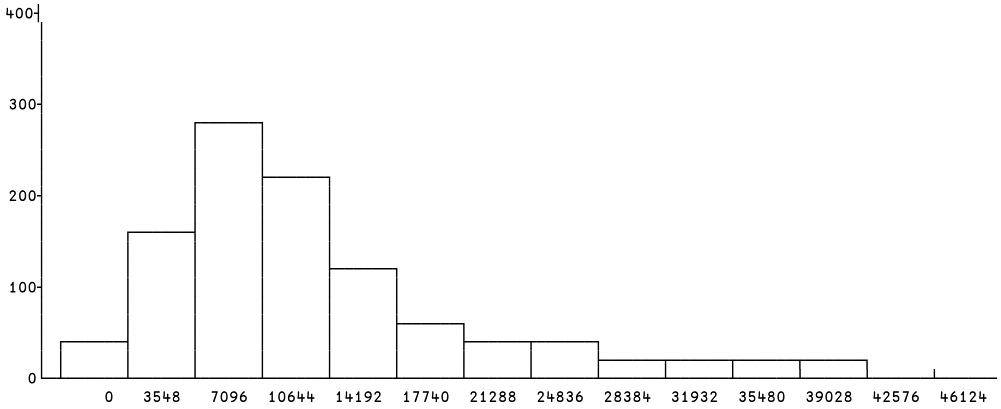
If the insurance company keeps $20,000 cash on hand to pay out claims each week, what is the probability that they don’t have enough money?

```
     proportion PAYOUT > 20000
0.147
```

Here we see that 14.7% of the time we will have insufficient funds to pay claims.

To get a better picture of the distribution of payouts, we will create a histogram which shows that the distribution skews to the right:

```
  histogram PAYOUT
```



The skewness measure of the distribution is positive indicating a right skew.

```
      skewness PAYOUT
1.538
```

We can also see that the distribution is right-skewed because the median is much lower than the mean.

```
     (median,mean)PAYOUT
9280.7 11564
```

7.5 Piedmont Airlines Problem

Piedmont Airlines uses a 19-seat aircraft for a flight between a small regional airport and a major hub at Boston's Logan Airport. The airline sells nonrefundable tickets for $150. There is a 10% probability that a ticketed passenger will not show up for the flight. For this reason, the airline will often overbook these flights. The cost for each bumped passenger is $325 which includes meals, an overnight stay in a hotel and lost goodwill. Management wants to know the optimal number of reservations to offer to increase profitability at the airline. Demand for seats follows the following distribution:

Demand(Seats)	14	15	16	17	18	19	20	21	22	23	24	25
Probability	3%	5%	7%	9%	11%	15%	18%	14%	8%	5%	3%	2%

First, we set up the probability distribution and generate a random variable for demand:

```
PROB←0.03 0.05 0.07 0.09 0.11 0.15 0.18 0.14 0.08 0.05 0.03 0.02
SEATS←14 15 16 17 18 19 20 21 22 23 24 25
DEMAND←SEATS PROB randomVariable 500
```

Now we can simulate the effect of overbooking by one seat (20 reservations):

```
SOLD←20 min DEMAND
```

We use the binomial distribution where the parameter n is itself a random variable (number of tickets sold). In that case the right argument to the random variable operator is not specified because we already know the number of random variables to generate.

```
BOARDING←SOLD 0.9 binomial randomVariable 0
BUMPED←0 max BOARDING-19
REVENUE←150×SOLD
EXPENSES←325×BUMPED
NET←REVENUE-EXPENSES
(mean,sdev) NET
```

Now we could run the above scenario over again for 21 all the way up to 25 reservations and pick the one with the largest average profit. However, we can automate the process by running everything simultaneously:

```
RESERVATIONS←19 20 21 22 23 24 25
SOLD←RESERVATIONS min enclose DEMAND
PARAMS← (enclose each SOLD), each 0.9
BOARDING←PARAMS binomial randomVariable each 0
BUMPED←0 max BOARDING-19
REVENUE←150×SOLD
EXPENSES←325×BUMPED
NET←REVENUE-EXPENSES
PROFIT←mean each NET
```

Now to find the highest profit:

```
I←PROFIT indexOf max PROFIT
RESERVATIONS[I]
scatterPlot PROFIT RESERVATIONS
```

7.6 Exercises

1. A regional airline uses a 25-seat aircraft for a flight between Scranton and Pittsburgh. The demand for this flight is binomial with n = 30, and p = 0.75. Tickets sell for $125. Expenses for this flight include the pilot's salary, fuel, and maintenance. Previous financial records show that the minimum expense per flight is $1000, the most likely expense is $1500, and the maximum expense is $3000. Find (a) the expected profit, (b) the standard deviation and (c) the probability the flight will lose money.

2. . A life insurance company sells a policy which pays $100,000 in case of death. The policy also pays hospital bills in case of injury. These bills average $50,000 with a standard deviation of $10,000. In any particular month there is a 5% chance of a death claim. Injury claims follow a Poisson distribution with an average of two claims per month. The company wants to estimate the total dollar amount of claims paid out each month to determine the minimum insurance premium it must charge customers to remain profitable. Run 1000 simulations and answer the following questions: (a) Find the expected value and standard deviation. (b) What is the probability that claims exceed $200,000 in any month? (c) What is the probability that they exceed $300,000 in any month?

3. Each day a newsstand purchases a certain quantity of papers. Demand is Poisson and averages 25 newspapers per day. The wholesale cost of newspapers is $1.00 apiece. If the newsvendor sells papers at $1.50 each, what is his average profit? What percentage of days does he lose money? How many papers should he purchase to maximize his gain?

Chapter 8 – Decision Analysis

The simplest decision is a choice between two options. If decision A has a payoff of $10.00 and decision B has a payoff of $5.00, one would naturally pick decision A. In real life decisions are made under uncertainty; thus if choosing between two stocks: Procter & Gamble (PG) and IBM, the return on a $1000 investment is:

Economy	PG	IBM
Recession	$30	-$50
Stable	$70	$30
Moderate	$100	$250
Boom	$150	$400

If we are optimists, we want to realize the greatest return assuming the best. This is called the **maximax** approach:

```
      ACTION←'PG,IBM'
      RECESSION←30,-50
      STABLE←70 30
      MODERATE←100 250
      BOOM←150 400
      PAYOFF_TABLE← ACTION RECESSION STABLE MODERATE BOOM
      NS←max decision max PAYOFF_TABLE
      NS.Values
150 400
      NS.Action
IBM
```

If we are pessimists, we want to realize the greatest return assuming the worst. This is the **maximin** approach:

```
      NS←max decision min PAYOFF_TABLE
      NS.Values
30 ¯50
      NS.Action
PG
```

In reality, the results will be somewhere in between the best-case and worst-case scenarios. If we assume the probabilities are equal, we can calculate the mean value:

```
      NS←max decision mean PAYOFF_TABLE
      NS.Values
87.5 157.5
      NS.Action
IBM
      NS.OptimalValue
157.5
```

Suppose there is a 10% probability of recession, a 40% probability of a stable economy, a 30% chance of moderate growth, and a 20% chance of a boom. What decision would we make now?

```
      PROBS←.1 .4 .3 .2
      NS←PROBS max decision mean PAYOFF_TABLE
      NS.Action
IBM
```

```
      NS.OptimalValue
162
```

We can often quantify the cost of uncertainty. How much would we be willing to pay if we knew the state of the economy ahead of time? This would be the expected value with perfect information:

```
      NS.EVwPI
186
```

The expected value of perfect information is the difference between the two:

```
      NS.EVPI
24
```

Although IBM produces the highest return on average, it is also riskier. If we wish to minimize our risk, we choose the stock with the smallest standard deviation which is Procter & Gamble:

```
      NS←PROBS min decision sdev PAYOFF_TABLE
      NS.Values
35.623 158.48
      NS.Action
PG
```

So there is a tradeoff between risk and return. We define the coefficient of variation as the standard deviation divided by the mean:

```
      CV ← sdev div mean
      NS← PROBS min decision CV PAYOFF_TABLE
      NS.Values
0.39146 0.97827
      NS.Action
PG
```

Another option is the return-to-risk ratio:

```
      RRR←mean div sdev
      NS←PROBS max decision RRR PAYOFF_TABLE
      NS.Values
2.5545 1.0222
      NS.Action
PG
```

8.1 Exercises

1. A tech company requires a certain computer chip and must decide whether to manufacture the chip itself or to purchase the chip from a vendor. Using the following table, which represents costs in thousands of dollars, and various strategies what decision would you make?

	Demand		
Decision	**Low**	**Moderate**	**High**
Manufacture Chip	750	1200	1950
Purchase from Vendor	1100	1750	1500

Appendix A – Reference Card

Function/Operator Syntax

X Y Z	Any array
A B	Boolean array {0,1}
C D	Character array
E	Nested Array
M N	Positive integer array
I J K	Integer array
P Q	Probability (0≤P≤1)
R	Unit (-1≤R≤1)
S T	Non-negative float (T≥0)
U V	Positive Float (U>0)
W	namespace
[X]	Optional
s/v/m/a	Highest rank allowed scalar/vector/matrix/array
f g h	functions
fD	Distribution Function
fS	Summary Function
fR	Relational Function
fL	Logical function

Operators

```
Pv ← (Xv fD)|Xv|Table prob fR Xv
Ps ← Cv1 (fL prob Table) Cv2
Ps ← Ps (fL prob ind)Qs
Xv ← Xv fD criticalValue fR Qv
Xv ← Xv fD randomVariable N
W ← [L]fD|Yv goodnessOfFit fR Xv
X1 X2← [P] fS confInt Xv
W ← [P] fS sampleSize T [U]
W ← Xv fS hypothesis fR Yv
XS ← Xv fS theoretical fD ''
W ← oneWay anova Xv Cv|Xv … Zv
W ← M [N] factor anova Xa
W ← [C] factorial2k anova Xa
W ← nested|blocked anova Xa
W ← [Cm] latinSquare anova Xm
W ← Bs leastSquares Yv Xv … Zv
W ← Yv regress [Bs] Xv … Zv
W ← Cv|f regress DB
W ← Yv ⊥ [N] regress [Bs] Xv [Zv]
W ← Yv f regress [Bs] Xv
Yv ← fS across Xm
Yv ← fS down Xm
Zm ← Yv f table Xv
Yv ← f|Cv pairwise Xv
```

Advanced Functions

```
NS ← Av independent Cv
Pv ← [Cv] bayes Pv Qv
```

Summary Functions

Measures of Quantity

```
Is ← count Xv
Zs ← sum Xv
Zs ← product Xv
Zs ← sumSquares Xv
Zs ← Yv sumProduct Xv
```

Measures of Center

```
Ys ← mean Xv
Ys ← median Xv
Ys ← mode Xv|Cv
Ps ← proportion Bv
```

Measures of Spread

```
Ts ← range Xv
Ts ← [Bs] var Xv|Cv
Ts ← [Bs] sdev  Xv|Cv
Ts ← iqr Xv
```

Measures of Position

```
Xs ← min Xv
Xs ← max Xv
X  ← I quartile Xv
X  ← P percentile Xv
P  ← Y percentileRank Xv
Z  ← Y zScore Xv
```

Measures of Shape

```
Zs ← skewness Xv
Zs ← kurtosis Xv
```

Measures of Association

```
Zs ← Yv cov Xv
Rs ← Yv corr Xv
```

Distribution Functions

Discrete Distributions

```
P←[N|1][Q|0.5] binomial I
P←Q geometric I
P←K N M hyperGeometric I
P←Cv|Iv Pv multinomial C|I
P←N Q negativeBinomial I
P←U poisson I
P←[[M|1] N] uniform N
```

Continuous Distributions

```
T←T U beta X
T←N chiSquare T
T←U exponential T
T←M N fDist X
T←T U gamma X
T←[X U | 0 1] normal X
T←X U logNormal T
T←X U rayleigh X
T←[X Y | 0 1] rectangular X
T←N tDist X
T←X Y Z triangular X [
T←X Y weibull T
```

Relational Functions

```
B ← X = Y  (A1 eq A2)
B ← X > Y  (X gt Y)
B ← X < Y  (X lt Y)
B ← X ≥ Y  (X ge Y)
B ← X ≤ Y  (X le Y)
B ← X ≠ Y  (A1 ne A2)
B ← X ∊ Y  (A1 in A2)
B ← A1 notIn A2
B ← X between Ys [Zs]
B ← X include Ys [Zs]
B ← X outside Ys [Zs]
```

Logical Functions

```
B2 ← not|~ B1
B3 ← B1 and|∩|∧ B2
B3 ← B1 or|∨| B2
B3 ← B1 given B2
B3 ← B1 nand B2
B3 ← B1 nor B2
B3 ← B1 without B2
B3 ← B1 iff B2
B3 ← B1 xor B2
```

Arithmetic Functions

```
T2 ← sqrt T1
Z ← ln U
U ← exp X
Z ← [Y] round X
Z ← sin|cos|tan X
Z ← arcsin|arccos|arctan X
```

Data Representation Functions

```
Z2 ← [U|-1]frequency  Xv|Av
W ← stats N M|(X S)
*W ← summarize  XX
Cv ← toNestedVector Av|Am
Am ← toMatrix Av|Cv
Av ← toDelimitedList Am|Cv
```

Display Functions/Operators

```
barChart C1v[C2v]
W←histogram Xv|Xm
W←pieChart Cv
Am←[U]stemAndLeaf Xv
Am←[U] dotPlot
W←[X]boxPlot Xv
W←Yv scatterPlot Xv
W←normProbPlot Xv
[Us] fS show Xv|Cv1[Cv2]
Am ← [Ps] report W
Z2 ← [U|-1]frequency  Xv|Av[Av]
```

Database Functions

```
W ← import AV.csv|''
Bs ← [Av] export W
W ← editDatabase W| Av
X ← [selectFrom] X where Y fR Z
W←[Cv] selectFrom W where Y fR Z
Yv Zv ← Xv splitBy Bv
```

Structural Functions

```
Zm  ←   [M N] matrix Xv
Zm  ←   transpose Xm
Zv  ←   N take Xv
Zv  ←   N drop Xv
Xm  ←   [M N] matrix Xv … Zv
Zv  ←   reverse Xv
Zv  ←   sort Xv
```

Utility Functions

```
NS  ←   load Av.dcf|''
NS  ←   save Av.dcf|''
NS  ←   import ''
Cm ← editTable  Cm | Av Av
SETUP ['R'|'DEMO'|'GUI'|'']
```